THE GOAL OF PHONETICS, ITS UNIFICATION AND APPLICATION

Björn Lindblom, Institute of Linguistics, Stockholm University,
S-106 91 Stockholm, Sweden

Chairpersons: Dennis B. Fry and Gunnar Fant

In trying to propose a formulation of the goals of phonetics I have begun by asking: (i) What are the goals and the methods of any scientific discipline? How does science in general work? secondly, (ii) What is the traditional subject matter of phonetics? and thirdly, (iii) What are some of the potential practical applications of phonetic knowledge?

Theory, explanation and scientific understanding

How do scientists formulate their understanding of the phenomena that they have chosen to investigate? We find generally that in empirical sciences it is in the form of a theory that such understanding is expressed. Consequently much scientific endeavor is directed towards the construction of theories. Accordingly a fundamental goal also of phonetics is theory construction.

Our first diagram (Fig. 1) is an attempt to illustrate in simplified form some of the components likely to be found in all scientific work such as making quantitative observations, deriving numerical predictions from a theory and inventing a theory. Scientists select a certain set of phenomena that they would like to explain. This set is the explananda in the right, empirical part of the diagram. They devise methods of observation whose output is intended to be facts not artefacts.

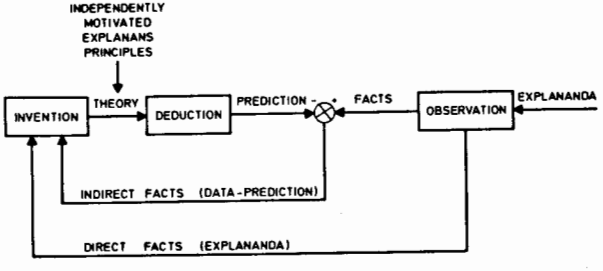


Fig. 1. Some components of scientific investigation.

Moving to the left we find the stage at which facts are compared with predictions or theoretical expectations. This is the point at which the evaluation of a theory begins. Or alternatively, if we have reason to be more confident in our theory than in our methods, it is the point at which we can assess the quality of our measurements. Early in my career as a phonetician I proudly showed Gunnar Fant some spectra that I had produced on the lab spectrograph with what I thought was extreme care so as not to introduce calibration errors etc. Much to my disappointment Gunnar dismissed the data right away and talked about distortion and "spurious formants". Of course he was right. But how could he tell? Later I have realized that the answer is that he looked at the data from the point of his strong theoretical understanding. I find this anecdote instructive since it pinpoints a general problem of research in the several areas of phonetics in which we still lack a powerful theory.

I shall use the term theory to refer to a set of basic laws or principles, on the one hand, and a system of rules on the other. From these basic principles and by means of these rules we deduce mathematically, in a perfectly automatic and formalized way, certain (numerical) consequences representing the predictions of the theory. The job that theories do is to explain. The anatomy of a scientific explanation presents at least the following parts:

1. It presupposes a theory that makes quantitative rather than qualitative statements.
2. It presupposes a theory that is completely formalized and leaves no room for the intelligence and intuition of the person using it.
3. It presupposes a set of explanans principles for which there is ample independent motivation. By independent motivation I mean justification not in terms of the data and the measurements but on external grounds.

In my usage the first two criteria are minimum requirements for an interpretation to qualify a theory. The quality of an explanation appears to be related to two things: the extent to which the theory meets the third condition, that is, has external justification and its scope, i.e., how much data it accomodates.

Summarizing what has been said so far we propose the following tentative definition of scientific understanding: To understand

something scientifically is to be able to recreate one's observations in a quantitative, formalized and explanatory way.

In order to further illustrate these ideas let us move back onto somewhat more familiar ground. Suppose we do an experiment in which listeners are asked to find the best perceptual match between steady-state pairs of synthetic vowels. The reference vowel has four formants. The test vowel has two. The upper formant, the so-called F_2', can be varied by the subject. Carlson, Fant and Granström (1970, 1975) did this type of experiment some time ago.

They were able to describe their results in two ways: (i) by means of an empirical formula making F_2' a function of F_2, F_3 and F_4; (ii) in terms of an auditory model reflecting the frequency analysis of the auditory periphery.

With respect to numerical accuracy the two descriptions gave almost identical and equally good results. However, when we place these accounts in the context of our previous discussion it becomes clear that only one of them offers an explanation, the one based on the auditory model. Why? Because this description is justified on external grounds. It shows us not only how but also why. It says that the matching behavior of the listeners is simply a consequence of a straightforward cognitive strategy and a phonetic universal: the human auditory system.

The empirical formula explains nothing. It captures certain regularities in the data in a compact and formalized way. It shows how the data came out but provides no clues as to why they came out that way.

Theory and explanation are concepts associated with the ultimate goals of research and it is therefore natural that most of the time we use these terms with restraint. We can name almost any area of phonetics: speech physiology, speech perception, speech development or sound change and we will find that in a certain sense it is true that "we are still at a data gathering stage". Note though that it would be a serious mistake to take this remark to mean that we should abandon all attempts at preliminary theoretical interpretation and model making and concentrate our efforts to the right half of Fig. 1. There are two types of data we need to gather: The facts obtained by direct observation, on the one hand, and the indirect facts represented by the discrepancies between the data and the theoretical model on the other.

Although the predictions may disagree with reality they should nevertheless be regarded as facts, facts about the model. Both the direct and the indirect facts are important sources of information in the creation of models. A good way to learn is to make mistakes in some systematic fashion.

The study of speech sounds: past and present

Phonetics has been traditionally defined as the study of speech sounds. If a deceased colleague of ours active around the turn of the century suddenly rose from the dead and could peep over the shoulders of his modern colleagues he would be unlikely to feel at home in our technologically sophisticated laboratories. However attending conferences and seminars he would no doubt conclude that the major problems to be solved and the questions asked had changed very little. It is instructive to contrast how classical phonetics dealt with the still current fundamental problem of devising a universal phonetic framework for spoken language. This task is essentially two-fold:

First of all, Find a way of describing phonetically an arbitrary utterance of an arbitrary language!

Secondly, Try to represent it in such a way that the description can be reproduced in audible form and with the linguistically relevant features preserved! Here the expression "linguistically relevant features" means the original native accent.

The first problem we can call the analysis or representation problem. The second is that of synthesis.

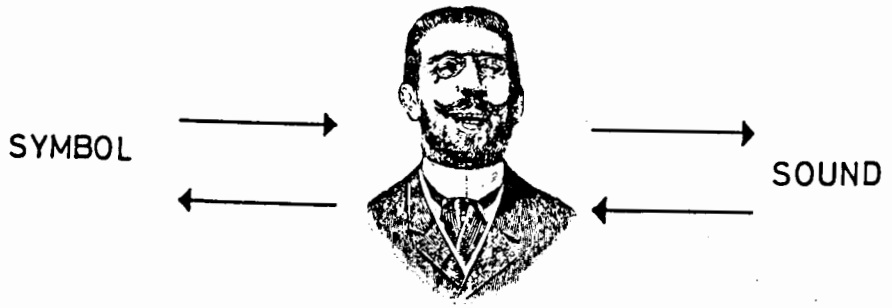


Fig. 2. The solution of classical auditory phonetics to the problem of speech sound specification: the skilled phonetician serving as a human tape-recorder in the "recording" and "playback" of acoustic facts.

The solution of classical phonetics was the concept of the universal phonetic alphabet and the use of highly skilled phoneticians serving as extremely sophisticated tape-recorders in the "recording" and "playback" of acoustic facts. Consider a certain utterance in a given language. Moving to the right in Fig. 2 corresponds to obtaining an answer to the question: What does this utterance, or rather the transcription of it, sound like? Moving to the left: The utterance just spoken by the informant, what is its representation in terms of phonetic symbols?

As we all know this solution of the problem of speech sound specification fails. Its inadequacies cannot be remedied by invoking the important insights contributed later by functional phonemic analysis and distinctive feature frameworks which achieved quantization of the infinite variety of sound and helped define the terms "alphabet" and "universal" more precisely. Nor would it matter if the quest for the ultimate phonetic framework could be brought to a successful close and if suddenly utopian phoneticians emerged capable of using transcription techniques of this type ideally. Why? If science aims at the construction of theories that explain the phenomena under investigation and if contemporary phonetics has the ambition to come of age as a science then it is quite clear why we reject the solution of classical auditory phonetics. This is so because the scientific description of speech sounds must necessarily aim at characterizing explicitly and quantitatively the acoustic events as well as the psychological and physiological processes that speakers and listeners use in generating and interpreting utterances. With the aid of the nimble tongue of the phonetic acrobat classical phonetics succeeds at best in skilfully merely imitating the speech processes of native speakers.

Clearly we must reject the method of impressionistic phonetics because it does not work in practice. Even if it did, it explains nothing: it does not reveal the processes underlying the production and perception of speech sounds. It does not represent a theory in the established sense of this term.

Phoneticians accordingly construe their task of speech sound specification as that of modeling the entire chain of speech behavior in a physiologically, physical and psychologically realistic manner. We thus arrive at the following conclusions: The

traditional subject matter of phonetics is the study of speech sounds; The general goal of scientific disciplines is theory construction and explanation; Consequently the goal of phonetics is to construct a theory of speech sounds; In order to make this theory meet established criteria of explanatory adequacy speech sounds cannot be studied as isolated acoustic events. Speech sounds can only be understood scientifically in terms of the psychological, physiological and physical processes responsible for their generation, on the one hand and with reference to their teleology, that is to their perceptual and communicative purpose on the other. Accordingly the phonetician whose inquiry began at the acoustic level in the domain of speech sounds is today forced to look upstream towards the mind and brain of the speaker and downstream towards the destination of the utterance in the brain and mind of the listener.

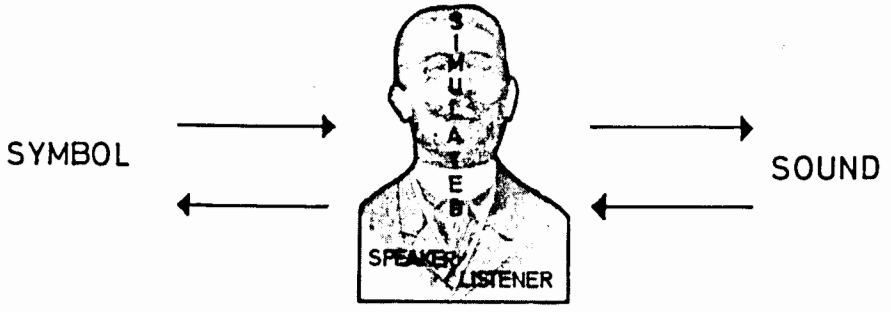


Fig. 3. A goal for modern experimental phonetics: a theory modeling the processes of speaking and listening in an acoustically, physiologically and psychologically realistic manner.

Let us at this point introduce Fig. 3, a slightly modified version of Fig. 2 and recall the phrase we used to summarize our initial discussion of scientific method: To understand something scientifically is to be able to recreate one's observations in a quantitative, formalized and explanatory way.

We can apply this thinking to a larger field of inquiry such as speech production, speech perception or speech development. Or we can apply it to a very restricted set of measurements made in a specific experiment. One very useful measure of our explicit rather than intuitive understanding of the phenomena investigated

is going to be our ability to recreate or simulate them. Needless to say we are in many cases not likely to come close to this goal in the foreseeable future. Nevertheless it provides us with the set of criteria we need to judge the relevance of our short-term efforts.

As we contrast past and present in the historical development of phonetics we see a discipline in the process of transforming from more or less a practical skill or an art into some sort of natural science. This development has yet to be completed but it is undoubtedly an inevitable consequence of: (i) the very nature of the subject matter that we have happened to have chosen; (ii) the natural ambition of any discipline to attain scientific maturity.

We should mention a third factor that has reinforced the present trend namely the prospect of using phonetics for practical purposes. Let me mention a few:

- educational methods and technical aids for the deaf, the hard of hearing, the handicapped and for second-language learners;
- the diagnosis and treatment of patients with phonetic symptoms including for instance delayed speech development, functional and organic voice disorders, aphasia, hypernasality, dysarthria and stuttering

as well as

- the automatic analysis and synthesis of speech for various technological purposes.

Békésy's mosaic model of scientific progress

In the introductory chapter of his book Experiments in Hearing, von Békésy describes his own research in relation to two research strategies: I quote "One, which may be called the theoretical approach, is to formulate the problem in relation to what is already known, to make predictions or extensions on the basis of accepted principles, and then to proceed to test these hypotheses experimentally. Another, which may be called the mosaic approach, takes each problem for itself with little reference to the field in which it lies, and seeks to discover relations and principles that hold within the circumscribed area." Further along in the text: "When in the field of science a great deal of progress has been made and most of the pertinent variables are known, a new problem may most readily be handled by trying to fit it into the

existing framework. When, however, the framework is uncertain and the number of variables is large the mosaic approach is much the easier. Many of the experiments to be described in this book employed the mosaic approach, but when considered in connection with other experiments carried out subsequently by the author and by many other workers in this field they take on a broader meaning and perhaps now may be woven into a more general structure."

Perhaps phonetics is a good example of a field growing like a mosaic. We have profited immensely from technological progress in the form of spectrographs, synthesizers and computers. Clearly such progress has not occurred as a result of premeditated planning on the part of phoneticians but as spin-off effects from adjacent fields with slightly different goals. Recruiting researchers trained in communication engineering, psychology, physiology, mathematics, physics etc. has demonstrably had an extremely vitalizing influence. According to the mosaic model of scientific progress the contents of a field is determined by the questions asked. Eventually a large number of questions will be asked and methods will be developed to answer them. Results will emerge that can be "woven into a more general structure". The lesson taught by the mosaic model thus seems to be: Leave your science alone! Stop worrying about where linguistics and phonetics are going and whether theoretical work is at a standstill or progresses sufficiently fast in response to practical needs etc. I would very much like to accept this advice. But unfortunately the examples that I am going to present to you will lead us in a different direction.

Form-based phonetics

When under laboratory conditions Swedish listeners hear the following stimulus:

Tape presentation of left spectrogram of Fig. 4 (next page).

Most of them say that they hear the Swedish word hallon beginning with an /h/ and meaning raspberry. What they hear and what you just listened to is in fact the following word simply played backwards[1]:

Tape presentation of right spectrogram of Fig. 4 (next page).

This word means zero. It has the so-called grave accent with an approximately symmetrical rising-falling F_0 contour. The spectrogram to the right thus shows the original recording and to the

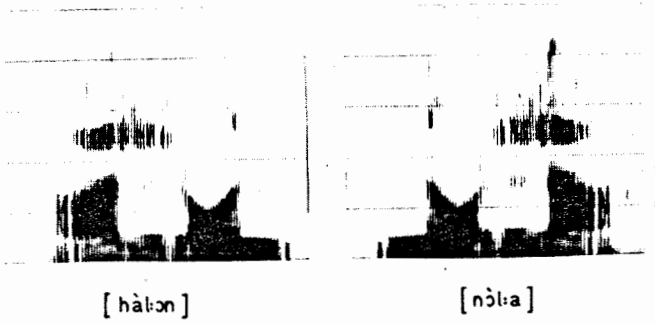


Fig. 4. A perceptual paradox: the "nolla-hallon" effect. Left: the Swedish word "nolla" played backwards. Right: the identical word played forwards. Transcriptions indicate perceptual asymmetry.

left we see the backward version. I think you can see that there is a weak expiratory [h]-like noise at the end of [nɔ̀l:a]. Why do our listeners perceive this segment as /h/ when we play the tape backwards but not forwards? One possible interpretation is that this perceptual asymmetry is due to the operation of top-down processes. In other words, you hear in terms of the structure of your native language. Like in many other languages the glottal fricative [h] does not occur in word-final or syllable-final position in Swedish. It does occur in initial position, however. Listeners do not have a sequence *allon, that is a sequence without the [h] in their lexicon. These facts evidently influence the perception of the acoustic signal in a drastic fashion for the effect is surprisingly strong to native Swedish ears.

The result of this simple tape reversal experiment appears to point to a fundamental principle of linguistic sound analysis: It is language structure and the human ear that determine what is linguistically relevant in the speech wave. The facts of physical phonetics cannot do so no matter how fine-grained we make the analysis. Although initially we rejected the method of classical auditory phonetics we are now paradoxically forced to admit that acoustic-instrumental facts about the behavior we are interested in must be accorded a secondary role in relation to the results of an auditory-functional analysis of sound substance. After all this is very elementary and not very new at all. Think of the notions of segmentation or invariance. Consider for instance the

distinctive feature, the phoneme, the syllable and so forth. All these are linguistic notions in the first place. They have an abstract theoretical status. We bring them with us into our laboratories (and normally we lose them in there before we get out).

Let us consider a statement by Malmberg (1968, 15). In the introduction of A Manual of Phonetics he formulates the role of experimental phonetics in a long-term perspective as follows: "...a combination of a strictly structural approach on the form level with an auditorily based description on the substance level will be the best basis for a scientific analysis of the expression when manifested as sounds. This description has to start by the fundamental analysis, then it must establish in auditory terms the distinctions used for separating phonemic units, and finally, by means of appropriate instruments, find out which acoustic and physiological events correspond to these different units. The interplay between the different sets of phenomena will probably for a long time remain a basic problem in phonetic research." Or take the following statement by Bolinger (1968, 13): "The science of phonetics, whose domain is the sounds of speech, is to linguistics what numismatics is to finance: it makes no difference to a financial transaction what alloys are used in a coin, and it makes no difference to the brain what bits of substance are used as triggers for language."

Substance-based phonology

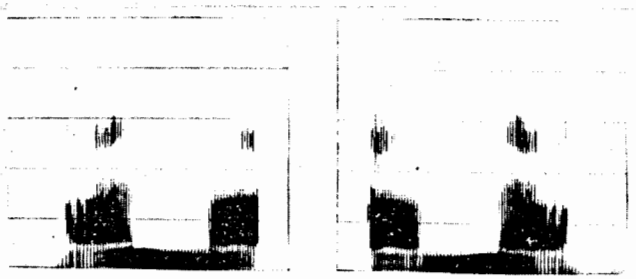


Fig. 5. Left: "Anna" (backwards). Right: the same word (forwards).

Investigating the case of syllable-final [h] further a colleague of mine at Stockholm University Eva Holmberg finds that this stimulus:

Tape presentation of left spectrogram of Fig. 5

is heard most often as Hanna. What you just heard was the following word played backwards:

Tape presentation of right spectrogram of Fig. 5.

Thus subjects clearly hear Hanna rather than Anna[2] in spite of the fact that both are names and should therefore be in the lexicon of our subjects. Clearly this throws some doubt on our previous interpretation attributing the perceptual asymmetry to language-specific top-down processing. A preliminary look at a large number of languages indicates that the /h/ phoneme tends to be either absent or realized as an [x]-laut or supraglottal fricative in syllable final position. These findings make us favor another hypothesis namely: The parallel between the perceptual asymmetry and the phonological asymmetry is not due to chance. It is due to universal properties of the human speech perception mechanism.[3]

The two cultures

The point that I would like to discuss is not whether this hypothesis is correct or not. Rather I have used the case of syllable-final /h/ to demonstrate that this hypothesis cannot be investigated within what Kuhn calls the current "paradigm" of linguistic theory. Given the role that phonetics has played so far in the construction of a theory of language there is no room for a hypothesis of this sort.

What is wrong? Although as linguists we are much concerned with the explanatory adequacy of our descriptions we nevertheless appear to make mistakes of a very elementary nature. In the beginning of our presentation we found that the concept of explanatory theory presupposes that reference is made to principles that are independent of the domain of the observations themselves and that have justification that goes beyond the patterning of the data (cf. vertical arrow at top left of Fig. 1). In common sense terms linguistic behavior presumably arises, both ontogenetically and phylogenetically, as the result of an interplay between

a) the functions that language is to subserve;
b) biological prerequisites such as brain, nervous system, speech organs, ear, memory mechanisms etc. and
c) environmental factors.

Languages thus evolve the way they do because of the body, the mind and the environment. They are the way they are on account of the functions they serve and owing to the properties of both innate and acquired mechanisms of learning, production and perception.

A scientific inquiry conducted along such lines would move our search for basic explanatory principles into the physics and physiology of the brain, nervous system and speech organs, the psychology of the mind and the social dimensions of language use. In other words it would take us right into areas that lie outside linguistics proper and the domains of our primary training and competence.

It might seem as if the strategy that I have been advocating is a reductionist approach to both phonetics and phonology. In other words, adopting this strategy would we then be headed ultimately for "molecular biology" rather than for insights of more primary interest to students of language? My response to this is that there are a host of phenomena for which we do not yet have a very good theoretical understanding. Just to mention a few consider the notions of distinctive feature, segmentation and the syllable and so forth. As long as we cannot treat for instance distinctive features as explananda, as things to be explained, rather than as empirically given primitives - as long as we cannot derive the distinctive features, that is the dimensions of possible phonological contrast, as consequences of constraints on speech communication the reductionist argument has very little force.

The history of phonetics and phonology is the story of two cultures that have always resisted unification. Trubetzkoy (Fischer-Jørgensen 1975, 22) classed phonetics among the natural sciences and assigned phonology to the humanities. The current paradigm of linguistics is aptly termed autonomous linguistics by Derwing (in press). In its context phonetics is a field worth annexing - but for completeness rather than for theoretical relevance.

One cannot help but suspect that autonomous linguistics and the role it assigns to phonetics has developed under the strong influence of educational and administrative constraints and that the program formulated by de Saussure and more recently by Chomsky is a brilliant rationalization of those constraints. If this suspicion is correct - and I truly believe it is - we have reason to examine how we train our linguistics and phonetics students and how without knowing it we become victims of the irrelevant and conservative influence of how universities are organized in terms of natural sciences, humanities and social sciences and so forth. In that kind of situation leaving one's science alone becomes impossible. However, educational programs can be changed.

Summary

We find that the long-term task of phonetics is to contribute towards the construction of a theory of language and language use. This goal is an ambitious undertaking calling for a multiplicity of experimental approaches as well as for theoretical unification.

The question of unification arises in all areas of our field but with particular force as we examine the traditional relationship between phonetics and phonology. We are forced to ask whether phonetics is currently embedded in an intellectual context that is ideally suited for approaching the long-term objectives. Generalizing from the results of a simple but I think instructive perceptual experiment I argue that the answer must be no. The trouble is that the stuff that theories and explanations are made of take us outside the domain of the primary training and competence of phoneticians and linguists. What can be done about this situation? Should we change the goals of phonetics? No, I don't think we can. We are trapped by our choice of subject matter, by scientific method as well as by our obligation to produce knowledge to fields of applied phonetics.

However, phoneticians are not alone in their dissatisfaction with the current paradigm of linguistics. Functionalism has always been alive. We see signs of linguistic research broadening its scope and intensifying research efforts in areas such as sociolinguistics, neurolinguistics, psycholinguistics, language acquisition, sound change, sign language, animal communication and so forth. I think this conference appears to demonstrate a number of such developments which inspire hopes for a new paradigm, a paradigm that views language in a biological perspective and makes it natural and respectable to ask teleological questions - questions that often successfully serve as guidelines for theoretical analysis in other areas of biology (Jacob 1970, Granit 1977) and that in the case of language patterns can be formulated as follows: For what biological and communicative purpose?[4)]

It seems to me that this is the paradigm that phonetic needs.

This is the paradigm in which phonetics will be most effective in contributing towards a better understanding of spoken language. That is a goal worth working for.

Conclusion

von Békésy found a close analogy to his research strategies in the field of art. To illustrate the mosaic approach he used a medieval Persian painting with persons and objects represented individually "with little perspective or relation to one another". For the theoretical approach he used a Renaissance woodcut constituting an early attempt to introduce perspective into representation.

Fig. 6. "The Gardener", painting by G. Arcimboldo (1527/30-1593), Skokloster Palace, Sweden.

I was inspired by Békésy to express my final point with the aid of a painting (Fig. 6). I would like to conclude by referring to a portrait by Arcimboldo (1527/30-1593). Let this painting be a symbol of three things: It symbolizes firstly the broad-based, multiple-approach experimental program that we should cultivate, secondly the need for theoretical unification and thirdly the hope that a biological perspective on speech and language will make such unification possible replacing the old paradigms of taxonomy and autonomous anti-functionalism.

Acknowledgments

The picture of the Arcimboldo painting was kindly made available to me by Svenska Porträttarkivet, Nationalmuseum, Stockholm.

Footnotes

1) The "nolla-hallon" effect was discovered about ten years ago by Ulf Ståhlhammar of RIT, Stockholm. I am grateful to him for bringing it to my attention at that time.
2) In working on the manuscript of this article I was pleased to hear from G. Heike that the "Anna-Hanna" asymmetry is valid also for German listeners.
3) The "nolla-hallon" effect resembles a phenomenon in psychoacoustics known as echo suppression. The sound of a hammer hitting a brick exhibits a certain decay waveform. Comparing backward and forward presentations of this noise one notes a striking asymmetry in that the decay appears much more prominent in the backward playback (Harvard Psychophysics Laboratory: Auditory Demonstration Tapes). There is some recent work on the forward and backward masking of speech-like noise stimuli caused by stationary vowels (Resnick, Weiss and Heinz 1979). This work shows forward masking to be more pronounced than backward masking. It is tempting to assume that the perceptual (and phonotactic?) asymmetry discussed here could be due to asymmetries of temporal masking among other things. However, the literature is somewhat ambiguous as to the direction and magnitude of these masking effects (Holmberg and Gibson 1979).
4) Note that I am not advocating some "divine foresight" responsible for order in nature. My model of "purpose" has two components: a "source" generating variation and a "filter" selecting those forms that happen to be compatible with certain "survival" criteria. In language communication the conditions of survival are social and biological in complex interaction.

References

Békésy, G. von (1960): "The problems of auditory research", in Experiments in hearing, 3-10, McGraw-Hill.

Bolinger, D. (1968): Aspects of language, New York: Hartcourt, Brace & World, Inc.

Carlson, R., B. Granström, and G. Fant (1970): "Some studies concerning perception of isolated vowel", STL-QPSR 2-3, 19-35.

Carlson, R., G. Fant, and B. Granström (1975): "Two-formant models, pitch and vowel perception", in Auditory analysis and perception of speech, G. Fant and M.A.A. Tatham (eds.), 55-82, London: Academic Press.

Derwing, B.L. (in press): "Against autonomous linguistics", in Evidence and argumentation in linguistics, T. Perry (ed.), New York: de Gruyter.

Fischer-Jørgensen, E. (1975): Trends in phonological theory, Copenhagen: Akademisk Forlag.

Granit, R. (1977): The purposive brain, Cambridge, Mass.: MIT Press.

Harvard University, Laboratory of Psychophysics (1978): Auditory demonstration tapes.

Holmberg, E., and A. Gibson (1979): "On the distribution of [h] in the languages of the world", PERILUS I, Dept. of Linguistics, Stockholm University.

Jacob, F. (1970): La logique du vivant, Paris: Gallimard.

Malmberg, B. (1968): "Linguistic bases of phonetics", in Manual of phonetics, B. Malmberg (ed.), 1-16, Amsterdam: North Holland.

Resnick, S.B., M.S. Weiss, and J.M. Heinz (1979): "Masking of filtered noise bursts by synthetic vowels", JASA 66, 674-673.

DISCUSSION

Victoria Fromkin, Hans Günther Tillmann, and Harry Hollien opened the discussion.

Victoria Fromkin: The question of the boundaries of phonetics and linguistics, or whether such boundaries should be drawn, is an important one. At the Linguistic Society Meeting in Salzburg last week, Charles Fillmore spoke on the question of boundaries, external and internal, in linguistics. The main point was that the goals we have are very often determined by which particular boundaries we set, and where we set them. And what is to one person phonetics may be to someone else garbage. It seems to me that we have to be able and willing to widen our boundaries.

When I first came into the field I was interested in electromyographic registrations of linguistic units, and there were people who said: "That is not linguistics", and I said: "But linguistics is whatever tells us more about the nature of human language and how language is realized in speech and in perception". - More recently I am interested in the human brain, and I am interested also in mental grammars, and I am even interested in what might go on in one part of the brain as opposed to another, - and people say to me: "That is not linguistics".

I have recently witnessed an experiment with a split-brain patient, whose left hemisphere, and subsequently right hemisphere were anaesthetized. When confronted with pictures of e.g. a mat and a bat, he could not tell them apart, - in fact, he could hardly speak at all, when his left hemisphere was anaesthetized. With the right hemisphere anaesthetized he did very well. Whether one has any quantitative results, whether there is one patient, ten patients, twenty patients, we know that there is something different going on in relation to when a person can tell the difference between mat and bat, and pig and big, and we do not even need to have more than five patients to know that there truly is something qualitatively different in the processing of the linguistic material from the non-linguistic, because when the left side of his brain was anaesthetized, this patient was still able to recognize and sing a song - so there is something special about the linguistic processing going on. Now, of course, we all know that, and what professor Lindblom did reveal is that to understand and to find explanations for this, we must go beyond our perhaps narrow interest and goals, and learn from the physicists, the neurophysiol-

ogists, the neurologists, the psychologists, and gain information wherever we can to try to understand both the nature of language as well as the way we use it in speaking and understanding. It is possible, and I think probable, that there will be certain aspects of human language which we will not find by just these kinds of research. And we will also learn that linguistic systems themselves will give us certain information, in fact raise certain questions as to what some of the rest of us in the laboratory have to seek answers to.

So where I agree with professor Lindblom is that we must go out of our own limited area, seeking help, information, explanations from various disciplines. But at the same time, I think that we should recognize that the autonomous linguists have some very important questions to raise for us to go and do our research on. I think that together we will begin to find out a little bit more about the intricate and complex nature of human language and about those of us who are users of it, the speakers, the hearers, and also the signers and perceivers of sign language, who are deaf.

Hans-Günther Tillmann: Professor Lindblom has drawn our attention to such fundamental and important problems as what it means to say that phoneticians try to develop theories which describe the phonetic facts of speech and language. To further clarify this issue, it could be helpful to turn to two somewhat simpler questions which, on this general metatheoretical level, are somewhat easier to answer: (1) What kinds of facts are given to the phonetician, and (2) what kinds of theories, according to the nature of these facts, can be developed by phoneticians?

(1) It is quite clear that all the facts that phoneticians are concerned with are given by concrete utterances produced by the speakers of a language. It is also quite clear that there are two different types of data to be found in these utterances. In natural circumstances, any such utterance can be perceived by a listener, say a trained phonetician, and hence it can be described symbolically. In this case, the phonetician's data are symbols, and he uses these symbols to refer to certain perceived (or perceivable) facts. Professor Lindblom gave us the two transcriptions [ana] and [hana], and everybody in the audience has learnt under which circumstances each of these transcriptions becomes

true or false - tertium non datur. Quite another type of fact comes into play as soon as we measure co-occurring variations in the physical world. These facts are transphenomenal to ordinary perception, at least in the case of phonetic variations co-occurring with perceivable utterances. If we measure these variations in different areas in and between the brains of the speaker and the listener ('signalphonetisches Band'), we obtain data in the form of time-functions, which in turn can be represented by digital signals. I would like to call special attention to the fact that these two different types of data, i.e. symbols and signals, constitute two different empirical domains for the phonetician - or, as I would like to call it if I could do so in English, two different 'empiries' - which exist separately and logically independently of each other. Perceivable utterances and measurable time-functions co-occur only empirically, yet in an experimentally reproduceable (i.e. verifiable) manner.

(2) Given these two different types of data - symbols, representing the category of perceivable events, and signals, representing measurable facts in the physical world - three different types of phonetic theories can be conceived of:

- A phonetic theory can be restricted to symbolic data - we find theories of this kind in phonology - or
- a phonetic theory can be more or less restricted to signals - the causal relations between different time-functions at different points of the physical continuum from the speaker to the listener can be analyzed in order to model the process of transmission of phonetic information from cortex to cortex - or
- a phonetic theory can explicitly try to connect the different facts given by symbols and signals - in this case, the form of a phonetic theory can simply be characterized by saying that the explicanda are primarily given in the first empirical domain of symbols, whereas independent explication can be looked for in the second empirical domain of time-functions.

Phoneticians and linguists are free to formulate and/or invent their explicanda, and they are also free to find theoretical explicata. In this situation, however, I would like to propose that phoneticians (and linguists) should make a virtue of necessity

and let practical applications determine what is to be translated into explicable explicanda. In this case, the solution of practical problems will be the best test to decide whether phoneticians have succeeded in finding a useful explicatum or not.

Harry Hollien: The first question that we have to ask ourselves seriously is: "Are we a discipline? Or are we simply a part of a more important discipline, whether it is linguistics or engineering or speech pathology, or some areas such as these?" - If we do decide that phonetics is a discipline, the second question we have to ask ourselves is: "Can we define it? Can we define its goals, its boundaries, its nature, in such a way that we can articulate this to other disciplines, and is there a cohesion within our field?" And since we represent different nationalities, different philosophies, different backgrounds, different orientations, different fractionalizations, we also have to deal with the third question: "How do we deal with each other, and develop mechanisms, procedures, processes by which to solve fundamentally the disagreements which we have within our field?"

Björn Lindblom: I think professor Hollien is doing it backwards. One begins by raising questions - that is how fields develop, that is how they grow. And: if phonetics is a discipline? I do not think it matters. We are interested in studying speech processes, interested in studying language, and that is where it all begins. And what you are talking about are some administrative, political problems that should be secondary.

I find myself in agreement with professor Fromkin and professor Tillmann. I wish that professor Fromkin would be a little more impatient with the autonomous linguistics paradigm, because it has such a radical influence on what we are doing.

Antti Sovijärvi gave examples of Finnish words which, when played backwards, are perceived by Finnish listeners in accordance with the syllable structure of Finnish.

Gunnar Fant pointed to the fact that there is a physiological explanation for the post-vocalic aspiration in open syllables, i.e. the glottis opens gradually, just like in an h-sound.

Henrik Birnbaum: In Björn Lindblom's initial chart (fig. 1) I was slightly disturbed by the terminology. He used the term 'indirect facts' for data prediction. But I do not think we can talk about fact in any sense here. We can talk, at best, about

hypotheses. I therefore do not think that there is any parallelism between the facts that we are asked to explain and the data predictions that we make, based on partial knowledge. - I think models are supposed to replicate something that we put into the abstract, and I think that what we have as 'indirect facts' in Björn Lindblom's chart, the data predictions, are part of a model, and models are never facts until they are proven beyond doubt correct, - so I would prefer the term hypotheses or partial hypotheses.

If 'autonomous' is understood in a broader way, and not in the narrow sense in which it was used in standard TG grammar, then of course autonomous linguistics, and within that autonomous phonetics, is a discipline. It does not mean that we should cut out all the neighbouring disciplines, however. I also would like to remind you that not only de Saussure and Chomsky would use this term, but Louis Hjelmslev spoke specifically about language as a structure sui generis. Language is a structure sui generis and not a replica of something else. We restructure reality in terms of the system we use.

Fred Peng: I want to ask professor Lindblom if he means that all people, regardless of linguistic or cultural background, hear more or less the same h initially, not heard at the end of the word. - Perceptual asymmetry is not limited to the auditory channel, it is also found in the visual and tactile modes, and I think that the environment, or context, has something to do with what you hear or do not hear, and the brain has sufficient plasticity to enable us to ignore what is not relevant to our background.

Björn Lindblom: We do not deny that listeners of different language background might have different perceptions, depending on their differences in top-down processing, conditioned by their native languages, but we do find parallels in the responses of our Swedish listeners and in the distribution of /h/-phonemes across the languages of the world. And thus we wonder if final /h/'s are not disfavoured because of some kind of auditory asymmetry that we all share. We are not denying that you can make use of this phoneme in final position, but it is disfavoured. It is a near universal absence.

Lise Menn: We need adequate descriptions from autonomous linguistics. It may well be that explanations cannot come from

within linguistics, but descriptions must. Early work in both child language and aphasia is, from a modern perspective, a great mess, - a lot of it, because of a lack of an adequate linguistic theory to relate the data to.

Another point: one level of investigation defines and sharpens the questions asked by another level. When you have gathered data for your theory, then you rephrase your questions - and it is a constant interaction between theory and data that is absolutely necessary. It is very easy to get a plethora of data: the problem is to relate it to theory. What you have is junk unless you know what its linguistic significance is.

Eric Keller pointed out the need for more theoretical papers in phonetics, the lack of which he tied up with the problem of educational background, which needs to be very wide if one is to do adequate work in phonetics. Students should be encouraged to acquire also mathematics, neurophysiology, physiology, psychology.

André Rigault suggested that we stick to de Saussure's distinction between substance and form: phonetics analyzes substance, phonology deals with form. He criticized the use of the term 'experimental phonetics' for something which is, properly speaking, 'instrumental phonetics', because doing an experiment involves having control of the phenomena investigated, to modify them at will. But he also felt that proper experimental phonetics ought to have a prominent place in our work, allowing us to verify theoretical models.

Further, phonetics should benefit from the contributions from psychology, linguistics, engineering, etc., but we should avoid the hyper-rationalization which has taken place in medicine, which produces people with a phenomenal education in mathematics, but no practicians to cure you of your illnesses.

Suzanne Romaine: I would like to object to the attitude which seems to be implied in professor Lindblom's last remark to the effect that a biological emphasis and perspective is what is needed to unify phonetics and to replace the old paradigms of taxonomy and autonomy, because it reflects a tacit acceptance of a Kuhnian notion of so-called normal science and of science as consisting of a succession of so-called paradigms. I think that unity is the last concept that should be applied to any discipline. We can agree about goals without having to agree on how we are

going to pursue them, and I would like to emphasize my agreement with what Victoria Fromkin said, that there are both quantitative and qualitative aspects to our profession. We do not want to be replacing old paradigms so much as to be increasing competition among paradigms. I think that is the only way for science to grow.

Pierre Divenyi: I would like to expand on the role of biology, from the point of view of perceptual phonetics, and say that maybe we should start learning from what our physiologist colleagues do: at the Cambridge meeting of the Acoustical Society of America in June, physiologists reported on experiments where they have measured the response to speech stimuli of various parts of the auditory system, and I think that now that we know at least how certain levels of this system respond, we should maybe cease considering as a stimulus to the phonetic system the string of phones, for instance, or even the acoustic stimulus itself. Maybe we should consider our proximal stimulus, to demonstrate what is happening at various parts of the system. I would tentatively suggest that the explanation for the 'Anna/Hanna' phenomenon shown to us by professor Lindblom may be deduced from what happens in the auditory nerve.

Fritz Winckel pointed to the parallel between natural sciences, linguistics, and art, all being trial and error processes.

Osamu Fujimura: The point I would like to raise is a general matter of how can we choose the correct criteria for selecting one model among several. And particularly, if there are two models at hand which both of them explain the facts equally well. We should probably be very careful about applying a particular set of criteria, because there are many cases where one experiment or situation does not reveal the entire picture of the subject-matter, and I think that for example in the case of the F2' experiments that professor Lindblom mentioned, isolated utterances, vowels, may not be revealing enough for us to be able to conclude in favour of one model over another.

Jørgen Rischel: It is obvious, to me at least, that we need autonomous linguistic research, at least a research which poses linguistic questions and which does not start out from, say, a biological foundation, and at the same time, of course, we need phonetic research. One of the problems today is that people specializing in different fields do not always grasp the implica-

tions of what people in other fields are doing. For example, it is very important to make clear to what extent a particular distinctive feature framework is motivated linguistically, to what extent it is phonetically motivated by, say, empirical physiological and perceptual research, and so on. There is sometimes a danger of a forth and back reinforcement of one's confidence in model construction: for example some linguistic model may serve as the basis for some phonetic experimentation and confirmation of the possibility of finding a phonetic equivalence, and then this may be used by the linguist as a confirmation of his own research. Therefore, we have to be very careful when we publish our results and make explicit whether we are borrowing assumptions which are not within our own paradigm or research.

REPORT: SPEECH PRODUCTION
(see vol. I, p. 11-56)

Reporter: Peter F. MacNeilage
Co-reporter: Peter Ladefoged
Co-reporter: Masayuki Sawashima

Chairpersons: Antti Sovijärvi and Hiroya Fujisaki

REPORTER'S ADDITIONAL REMARKS

P. MacNeilage, in his presentation, commented on the question of the control of speech production and the biological basis of speech.

The first comments dealt with the role of feed-back. P. MacNeilage claimed that if one considers how we produce speech under the various postural circumstances, we are forced to conclude that peripheral somatic feed-back plays a virtually continuous role in the control of speech production. It must be a system that can sense at the periphery what the present posture is and that is required to monitor the attempts of the control system to produce speech in any particular posture. We can't be assumed to be infinitely versatile in terms of preprogramming at all postural circumstances. Furthermore, P. MacNeilage pointed out that the concept of normal speech production is perhaps misleading, since most of our work is done in the laboratory with the subjects looking straight ahead and in a fixed position. This is not the normal posture and very little of our work has dealt with postural variations. P. MacNeilage continued by saying that we know very little about how the feed-back works and that we need more information which may perhaps come from people doing research in dentistry. He warned against conclusions drawn from physiological studies of animal limbs, since the human somatic sensory system differs from the animal system in many significant ways. In addition, P. MacNeilage found that the results of experiments where the posture is artificially manipulated, such as in the bite-block studies and in studies where the jaw movement is impaired, support the argument about the necessity for feed-back.

Then P. MacNeilage raised the question: This feed-back is feed-back to what? Among other things he pointed out that it seems necessary in speech production to recognize a multiplicity of levels of organization, some of which are quite accessible to us

and others which are not. But it is nevertheless crucial for us to understand those higher levels if we want to come up with a plausible theory of speech production. In this connection, P. MacNeilage stated that there has to be a distinction between a context sensitive system at a lower level of organization and some kind of context independent entity or set of entities at a higher level, referring among other things to segmental spoonerism. We produce sequences with spoonerisms fluently which means that subsequent to the permutation, the context sensitive control system makes the appropriate adjustments. He noticed that very often spoonerisms involve single segments, and very few can be unequivocally labeled distinctive feature movement type errors, and relatively few involve whole syllables. This means that at least at one level of organization the segmental unit is an extremely important one for speech production.

Before leaving the topic of control, MacNeilage stated that our rather simple algorithms do not account very well for the dynamic aspects of speech production, referring to differences in stress and speaking rate, and to coarticulation. The same speaker can use different strategies in changing the speaking rate, for instance, which also proves that we are dealing with an extremely versatile control system.

Turning now to the question of the biological basis of speech production, P. MacNeilage emphasized - as he does in his paper - that we have very much neglected the study of prelinguistic vocalization in our studies of speech production. This neglect may be due to R. Jakobson's theory of language acquisition which assigned babbling to "external" phonetics. P. MacNeilage claimed that the phonetic forms of early speech with reference are extremely similar or identical in many cases to the babbling forms that immediately preceded them. This means that the same production system that has been working earlier in the proto-language stage is still an extremely important component in early referential speech. P. MacNeilage claimed that babbling begins at a particular time on a particular day. Finally, he stated that babbling is some kind of innate movement control organization that is "there" in relation to speech.

DISCUSSION

John Ohala and John Laver opened the discussion.

J. Ohala stated that from his point of view one of the very promising and most essential developments in current work on speech production is the large number of models, including various aspects of the articulatory apparatus, which have been developed in the past decade or so. He believes that the rise of model-making is a development of the computer revolution in the laboratory and that it has come of age where we have become familiar with and have used computers to develop models which in many cases are conceptually simple, but which require computationally rather complex activity. Some objections have occasionally been raised against model-making, usually along the lines of: "Well, you have made the model, you have put the properties into it that it has, why can't you figure out what it is supposed to do in advance, why bother with it? It is simply making explicit what you already know or what you assume to be true." In order to parry off this kind of objection, Ohala referred to the Nobel Prize winner H. Simon, who indicates that it may very well be true that in model-making-like abstract logic and didactic logic and so on - the consequences of a particular set of assumptions must naturally follow in an automatic, perfectly regular way. But when our models and the assumptions in them get sufficiently complex, really only God can figure out what the consequences of these assumptions may be. The rest of us have to work them out painstakingly, teasing them out for understanding, and this is why we make models. Furthermore, Ohala pointed out that our models serve a very interesting heuristic purpose in that they tell us what to look for in the data. This was made evident to him in working with an aerodynamic model revealing that if one is going to have production of a fricative or some kind of fricated segment one should not have nasal leakage, obviously because the air flowing out of the nasal cavity would prevent the build-up of the high pressure drop necessary to produce the turbulence. And Ohala asked whether this has phonological consequences. He pointed out that he had never seen any observation of this in the literature, but when he searched for it he was able to come up with a number of examples from sound change and allophonic variation. For example, English has a palatal fricative

as an allophone of /h/ before the palatal glide /j/ in words like Hugh and human. But that same allophone is no longer a fricative if we embed it in a heavily nasalized environment as in the word inhuman. With this example Ohala illustrates how models can tell us what to look for and in that sense even help us to enhance our naturalistically obtained data base.

Then Ohala addressed one comment to Sawashima concerning the vertical tension of the vocal folds. Sawashima said in his co-report that there is no evidence for the existence of any physiological mechanism whereby vertical compression or tensing of the cords could affect F_0. However, it is well known that the average F_0 of vowels is positively correlated with the "height" of vowels. But, to date, no one has found any significant difference in the degree of muscle activity of the intrinsic laryngeal muscles during the production of various vowels. On the other hand, van den Berg (1955), Shimizu (1960, 1961), and additional workers cited in Žinkin (1968:353) have found that the laryngeal ventricle is larger, both in width and vertical depth, during the production of high vowels such as [i] and [u] - thus showing greater separation between the ventricular folds and the vocal folds - but smaller during the production of low vowels. Also, Luchsinger and Arnold (1965:223) describe a patient with bilateral paralysis of the cricothyroid muscles but who could nevertheless vary F_0 over a few semitones. X-rays revealed no change in the angle of the cricothyroid visor but the whole larynx was higher in the neck during the production of high F_0. (More detailed arguments for F_0 variation due to vertical tension have been given in Ohala 1972, 1977, 1978.)

References

Berg, J. van den (1955): "On the role of the laryngeal ventricle in voice production", Fol. phon. 7, 57-69.

Luchsinger, R. and G.E. Arnold (1965): Voice-speech-language; clinical communicology: its physiology and pathology. Belmont: Wadsworth.

Ohala, J. (1972): "How is pitch lowered?", JASA 52, 124.

Ohala, J. (1977): "Speculations on pitch regulation", Phonetica 34, 310-312.

Ohala, J. (1978): "The production of tone", in Tone: a linguistic survey , V. Fromkin (ed.), 5-39. New York: Academic Press.

Shimizu, K. (1960): "On the motions of the vocal cords in phonation studied by means of the high voltage radiograph movies". [In Japanese; English summary]. Oto-Rhino-Lar. Clinic [Zibi-Inko-Ka Rinsyo] 53, 446-461.

Shimizu, K. (1961): "Experimental studies on movements of the vocal cords during phonation by high voltage radiograph motion pictures", Studia Phonologica 1, 111-116.

Žinkin, N.I. (1968): Mechanisms of speech, The Hague: Mouton.

J. Laver had four points to make about the issues raised in the three reports.

The first of them dealt with methods for estimating the different muscular forces acting on and in the tongue, as in the work of Fujimura, Kakita, and Perkell. He referred to a finding from speech error work based on an experiment to provoke subjects into making the kind of vowel-blend errors that Rulon Wells claimed almost never happen. The structure of the experiment was to push subjects just beyond the comfortable limit of accurate performance of target vowels. Facing them on a screen were two words - for example PEEP and PARP - and above the two words were two stimulus lights, and the task was to pronounce each word as accurately as possible immediately the associated light came on. The lights were programmed to come in random sequence, with 200 msec duration, with intervals of 200 msec. In this condition, all subjects made vowel errors, two types of diphthongs and one type of monophthong. When PEEP and PARP were in competition the two diphthong errors were either PAIP or PIAP. Laver proposed the following hypothesis to explain this result. One might imagine that the commands to the relevant muscle systems had a slight difference in the time course such that if the commands for AR preceded those for EE then one got PAIP and if the commands for EE preceded those for AR one got PIAP. But if the commands to the different muscle systems were issued perfectly simultaneously, then the monophthong [ɜ] as in PURP was the result as the mechanically joint product of the action of simultaneously activated different muscle systems. The relevance of this finding to the problem of estimating relative muscle system forces is, that if we look at the interactions of all pairs of vowels, then the "mechanically joint product" position of the intermediate vowel does not necessarily coincide with the geometric mean position between the two target positions. In the competition between PEP and POOP, for example, the intermediate monophthong was [œ] as in [pœ:p], in other words rather closer to the [ɛ] target than to the [u] target, as the lip position also, one might think, was slightly closer to the [ɛ] target

than to the close rounded [u] target. And this is, as far as the tongue is concerned, presumably because the genioglossus muscle has greater muscular force than the muscle system that raises and backs the tongue. Muscle system interactions of this sort in the balanced protagonist-antagonist situation in ordinary speech may well lie at the basis of the notion of "favoured articulatory zones" in the languages of the world. Laver concluded that we have here a very simple experimental paradigm of competition between two targets programmed in random sequence at high speed which can be applied in many areas of speech production and which can tell us perhaps a number of interesting things about the way speech is represented and controlled neuromuscularly.

Secondly, Laver had a comment about Ladefoged's suggested laryngeal parameters of glottal aperture, glottal tension, and glottal length. He pointed out that one aspect of the usefulness of this approach is that the six main modes of phonation - modal voice (Hollien's term), falsetto, creak, whisper, breathiness, and harshness - all have different specifications on these three parameters. And therefore, an explanatory basis is provided for the mutual compatibility or incompatibility between these six phonatory modes. It means that breathiness and harshness, for instance, are ruled out by that model as mutually incompatible, as they are in real life, because they need very different values on the glottal aperture and the glottal tension parameters.

Laver's third point concerned the habitual mode of phonation adopted by an individual which he found was an excellent example of a muscular setting (Honikman's term). The notion of a setting is extendable beyond the larynx to habitual adjustments of the supralaryngeal tract as well. We are all familiar with people using a particular long-term muscular adjustment of the supralaryngeal tract as part of their habitual voice quality. For example people who raise their larynxes and keep them raised throughout speech, people who have a tendency to maintain the lips protruded, qualities which characterize particular speech communities like velarization that one hears in the speech of Liverpool, and lastly habitual nasalization common among RP-speakers. The nice thing about muscular settings, in the context of MacNeilage's report, is that they furnish an excellent example of the Action Theory concept of co-ordinative structures, tuned to a long-term bias on segmental articulation - just like habitual gait.

The last point dealt with the problem of neuromuscular programming, when it is not just a matter of programming a sequence of segments as such, but rather of programming at least a triple layer of commands. Laver stated that if voice quality has a phonetic component which demands a particular controllable setting of the vocal tract and the larynx, then one has to take care of the neuromuscular programming for that component. Secondly, superimposed on that phonetic component of voice quality there will be the current tone of voice that the person is using, in other words the paralinguistic layer as well. And thirdly, the segmental and other components of the linguistic strand of speech. Laver concluded that neurolinguistic programming in real speech is at least three times more complex than would be needed for any single-layer control of segmental sequence.

M. Sawashima, responding to Ohala's last point, claimed that he did agree that the up and down movement of the larynx is highly correlated to the F_0 change. But Sawashima found it difficult to explain that the up and down movement of the larynx directly can affect the vocal fold tenseness if we consider the mechanical and structural properties of the larynx. Maybe we can explain it by saying that the up and down movement of the larynx indirectly can provide a change in the longitudinal tension of the vocal folds, which was said many years ago by Sonninen and others. Sawashima concluded that what we want to find is a reliable physiological correlate to the change or control of the vocal fold tension, and in that sense we can't say that the change of the vocal fold tension is caused by the up and down movement of the larynx.

S. Smith drew the attention of the audience to some of his works dealing with the functional dichotomy of the vocal folds (membrane-cushion, cover-body) and which were done before the works made by van Berg and Hirano.

P. Ladefoged presented a series of slides showing the laryngeal behaviour for different voice qualities in a Bushman language. In his co-report Ladefoged pointed out that the laryngeal parameters normally used are completely inadequate for a description of the six voice qualities found in this language. A very interesting finding was that the speakers of the language all had

a thickened interarytenoid muscle, which helps them to produce the ventricular phonation. The bulge seen on the interarytenoid muscle is not genetically controlled, because one of Ladefoged's colleagues has developed a thickened interarytenoid muscle, working with the language.

O. Fujimura had two points to make. The first one dealt with spoonerisms as evidence for the phoneme size segment as the functional unit. He pointed out that no unit whether phoneme, distinctive feature, syllable, or word can freely exchange with another unit in any environment. The facts are more complicated, and there are constraints and contextual conditions that have to be considered. Fujimura found that there is a confusion between the elements for exchange and the environment set up for the exchange of the elements, and he proposed to consider not only one unit for everything, the phoneme for instance, but also larger units as well. Typically, the exchange occurs in syllable initial position, and why is it so if the phoneme is really the functional unit for exchange?

The other comment concerned the vertical movement of the larynx, which Fujimura found is a very interesting phenomenological fact in correlation with pitch control. This is quite useful in finding out what the control signals are for "pitch control" in devoiced portions of speech. He referred to Japanese which has vowel devoicing according to certain contextual conditions. Fiberoptic observations have shown some vertical movement, qualitatively, in relation to the lexical accentual patterns and also to the phrase boundary phenomenon. In the case where the second syllable of the phrase is devoiced and should be high in pitch according to the general rules, some native speakers feel that the second syllable in those devoiced cases is low in pitch. And fiberoptic observations seem to support this feeling in terms of the vertical movement of the larynx.

N. Waterson had some comments concerning the question of babbling as preparation for speech. Early babbling or cooing usually begins spontaneously as a type of unstructured vocalization and is generally mainly vocalic in nature with perhaps a few sounds in the velar and uvular regions. This stage seems to be

non language-specific. But Waterson pointed out that the interactions between the baby and his caretakers play an important role in preparing the baby for linguistic communication.

The vocalic type of vocalization is replaced by more complex vocalizations containing various consonantal sounds, and they become structured and repetitive. This suggests that the baby is developing processing skills which enable him to recognize samenesses and differences in vocal stretches - something that is essential for the development of language. When structured babbling begins, mothers tend to imitate those stretches which seem to them to be similar to their own language, so the baby is encouraged to work on the sounds of the language of the environment. The child is thus prepared for the sounds he will use in his first words.

The protolanguage stage, which usually overlaps with babbling, is generally articulatorily much less complex than what has been achieved in babbling but represents the development of the functional use of vocalizations. When vocalization is first used functionally, the production is very simple as if articulatory complex production and functional use cannot be coped with by the child's processing system at the same time at this early stage. When he has learnt how to use simple vocalizations functionally, he is ready for the use of the more complex production of the actual speech, and the first words soon follow.

B. Lindblom pointed out that the interest in the biological basis of speech, brought up by MacNeilage, is an interest in the most general phonological universal of all, namely in the difficult topic of speech sounds being a subclass of all sounds and gestures. In this context Lindblom had a question for Ladefoged, Fujimura, and Perkell, which had to do with our articulatory modelling: "Why leave out the jaw?" Lindblom had earlier argued that with the aid of the notion of neutral tongue shape and the jaw parameter we can perhaps explain the origin of the distinctive feature open and close. Furthermore, Lindblom referred to some jaw data presented in his symposium report showing how consonants resist coarticulation in the environment of maximally open vowels. He found that this illuminates some of the phonetic background on phonotactic syllable structure, on strength hierarchies, and such abstract notions from phonology.

J.S. Perkell mentioned, responding to Lindblom, that the actual contribution of the structure of the jaw - i.e. the lower teeth - to directly determining the area function is minimal, but that the jaw serves more as a framework for carrying other articulators around and thereby has an indirect influence. Perkell pointed out that we can't answer the question concerning the importance of the jaw without including the jaw in our physiological models.

P. MacNeilage, replying to Fujimura, mentioned that what he really wanted to say was to stress the prominence of the segment assuming that the larger the number of areas that involve a unit the more important it is at a particular stage of the modelling process to which one thinks the areas are relevant. He agreed that one has to take into account many units in the modelling process and that contextual influences are extremely important.

Replying to Waterson, MacNeilage pointed out that by babbling he did not mean cooing but just what he liked to call the canonical form, the open-close alternation with time locking. He found that maybe he disagreed with Waterson about the onset of that stage. MacNeilage was of the opinion that it happens rather suddenly. It is an important point that has to be explored in the light of the role of imitation. If the adults imitate the child's forms but the child's initial forms occur suddenly, then imitation may have a rather minor role in the onset of the phenomenon, even if it may be important it its subsequent development.

MacNeilage concluded by saying that he was impressed with the lack of disagreement that there had been about the speech production aspect of the phonetic discipline. He liked to believe that it is a very healthy sign and that the heat of the argument is related to the state of the knowledge in this area.

P. Ladefoged returned to the problem dealing with the jaw. His evidence to say that one should leave out the jaw is that what is controlled is the vocal tract shape, referring to Lindblom and his colleagues, who have shown quite effectively that we can produce very similar shapes with the jaw in different positions. If we look at mathematical techniques for reducing the amount of variance between a group of speakers we come out with factors that

reflect the cavity shapes and do not reflect the jaw positions. This is another evidence that the jaw has no role to play. But Ladefoged pointed out that it is just so for vowels and that he might have to put the jaw back again for consonants, referring to Lindblom's new jaw data for consonants (cf. vol. II, p. 33-40).

H. Fujisaki mentioned that we have to treat the jaw as an independent motor unit when we are dealing with the dynamics of articulation. When Ladefoged speaks about tongue control it is a combination of independent or dependent control of the jaw and the tongue. The fact that one can produce many speech sounds without moving the jaw does not exclude the fact that the jaw plays an important role in articulation.

N. Waterson, responding to MacNeilage, replied that if he by "sudden" meant over two or three weeks then there was probably no disagreement, but if he meant from one day to another then they did disagree. But she pointed out that there is not quite enough data on babbling to be able to make a categorical statement about it.

MacNeilage admitted that he did not have very much data and that much of it was informal, but it was his impression that it happens virtually from one day to the next.

Fujimura advocated the independent function of the jaw. He referred to his tongue model, which actually includes an independent variable corresponding to the jaw angle. Fujimura found that the jaw has important functions particularly with respect to the inflection of stress patterns referring to some of his jaw data, which show that jaw height does not correlate clearly either positively or negatively with tongue height and it is not random either. He concluded that the jaw constitutes a very important articulatory dimension.

J. Ohala made a comment dealing with the interpretation of speech errors. He did agree with Fujimura's call for caution in the interpretation of speech error data for what they may reveal about units of speech production. He did this with an analogy.

Let us imagine the following domestic accident: a cook stores spices in a spice cabinet in alphabetic order, i.e., cumin is after corriander, and tumeric is after thyme, etc. In reaching for the thyme to add to a dish, the cook accidentally grabs tumeric instead, thus making a culinary analogue of a speech error. The analyst trying to interpret this error would look in vain for any chemical or physical similarity between tumeric and thyme. What is the point of this? Simply that for the purpose of retrieval or general "housekeeping" functions of manipulating the stored units of speech, it is possible that the addresses or labels used bear only an arbitrary relationship to the substance of the units themselves. Ohala concluded that until we have some general idea of how speech is "programmed" he did not think that the data from speech errors can unambiguously rule out features, phonemes, or syllables - or something else - as possible units of production.

REPORT: SPEECH PERCEPTION
(see vol. I, p. 59-99)

Reporter: Michael Studdert-Kennedy
Co-reporter: Hiroya Fujisaki
Co-reporter: Ludmilla Chistovich

Chairpersons: Antony Cohen and Louis C.W. Pols

REPORTERS' ADDITIONAL REMARKS

Michael Studdert-Kennedy gave a summary of his report. He mentioned that he might have misunderstood the aim of the work of the Leningrad group to some extent. He had thought that they were looking for phonetic segments in the acoustic signal, i.e. for acoustic segments that would be isomorphic with phonetic segments, but it appears from Ludmilla Chistovich's report that they are in fact looking primarily for acoustic segmentation, which will, e.g. be essential for the estimation of durational events.

Discussing the problem of feature detectors he mentioned that animals that have feature detectors and templates (e.g. the bullfrog and birds) have them because they need them, having to get along very soon after birth without parental help, but that is not the case with the human infant, who has a long period of parental care.

Concerning the problem of perception of sounds by means of an integration of a variety of cues, he emphasized that the idea that these cues may be held together by the underlying gesture should not be understood as a claim for a motor theory of perception, which implies that perception requires reference to the production system. The idea is that you perceive the production gesture directly like you perceive the movement of a hand by means of the light reflected from it. If the hand was moved inside a resonating chamber which had a source exciting it, you might hear the gesture instead of seeing it.

Studdert-Kennedy added a section on cerebral specialization not found in the original report. A written version of this addition is given below:

Cerebral specialization

Nonetheless, opposition between the two modes of lexical

access -- holistic, from "auditory contour", analytic, from phonetic segments -- should not be too sharply drawn. The work of Zaidel (1978a,b) with "split-brain" patients has demonstrated that holistic access is certainly possible. The cerebral hemispheres of such patients have been surgically separated by section of the connecting pathways (corpus callosum) for relief of epileptic seizure. The separation permits an investigator to assess the linguistic capacities of each hemisphere independently. Zaidel (1978 a,b) has shown that the isolated right hemisphere of such a patient, though totally mute, can recognize a sizeable auditory lexicon and has a rudimentary syntax sufficient for understanding phrases of up to three or four words in length. However, it is incapable of identifying nonsense syllables or of performing tasks that call for phonetic analysis, such as recognizing rhyme (cf. Levy, 1974). This phonetic deficit evidently precludes short-term verbal store, thus limiting the right hemisphere's capacity for syntactic analysis of lengthy utterances, and forces organization of language around meaning. Whether we assume a similar, subsidiary organization in the left hemisphere or some process of inter-hemispheric collaboration, it is clear that normal language comprehension could, at least in principle, draw on both holistic and analytic mechanisms.

At the same time, Zaidel's work provides striking support for the hypothesis, originally derived from dichotic studies, that the distinctive linguistic capacity of the left hemisphere is for phonological analysis of auditory pattern (Studdert-Kennedy and Shankweiler, 1970). Further support has come from electroencephalography (Wood, 1975) and, quite recently, from studies of the effects of electrical stimulation during craniotomy (Ojemann and Mateer, 1979). The latter work isolated, in four patients, left frontal, temporal and parietal sites, surrounding the final cortical motor pathway for speech, in which stimulation blocked both sequencing of oro-facial movements and phoneme identification.

This fascinating discovery meshes neatly with a growing body of data and theory that has sought, in recent years, to explain the well-known link between lateralizations for hand control and speech. Semmes (1968) offered a first account of the association by arguing, from a lengthy series of gunshot lesions, that the left hemisphere is focally organized for fine motor control, the right hemisphere diffusely organized for broader control. Subsequently,

Kimura and her associates reported that skilled manual movements (Kimura and Archibald, 1974) and non-verbal oral movements (Mateer and Kimura, 1977) tend to be impaired in cases of non-fluent aphasia. These impairments are specifically for the sequencing of fine motor movements and are consistent with other behavioral evidence that motor control of the hands and of the speech apparatus is vested in related neural centers (Kinsbourne and Hicks, 1979). In fact, Kimura (1976) has proposed that "...the left hemisphere is particularly well adapted, not for symbolic function per se, but for the execution of some categories of motor activity which happened to lend themselves readily to communication" (p. 154). Among these categories we must, incidentally, include those that support the complex "phonological" and morphological processes of manual sign languages, now being discovered by the research of Klima, Bellugi and their colleagues (Klima and Bellugi, 1979).

The drift of all this work is toward a view of the left cerebral hemisphere as the locus of interrelated sensorimotor centers, essential to the development of language, whether spoken or signed. To understanding of the speech sensorimotor system perceptual studies of dichotic listening will doubtless contribute. Indeed, important dichotic studies have recently found evidence for the double dissociation of left and right hemisphere, speech and music, in infants as young as two or three months (Entus, 1977; Glanville, Best and Levenson, 1977). However, dichotic work has not fulfilled its early promise, largely because it has proved extraordinarily difficult to partial out the complex of factors, behavioral and neurological, that determine the degree of observed ear advantage (cf. Studdert-Kennedy, 1975). For the future, we may increasingly rely on instrumental techniques for monitoring brain activity, such as the blood-flow studies of Lassen and his colleagues (Lassen, Ingvar and Skinhøj, 1978), induced reversible lesions by focal cooling (Zaidel, 1978b), improved methods of electroencephalographic analysis, auditory evoked potentials (Molfese, Freeman and Palermo, 1975) and, perhaps infrequently, direct brain stimulation.

References

Abramson, A.S. (1977): "Laryngeal timing in consonant distinctions", Phonetica 34, 295-303.

Campbell, R. and B. Dodd (in press): "Hearing by eye", Quarterly Journal of Experimental Psychology.

Entus, A.K. (1977): "Hemispheric asymmetry in processing dichotically presented speech and nonspeech stimuli by infants", in S.J. Segalowitz and F.A. Greber (eds.) Language development and neurological theory, 64-73, New York: Academic Press.

Glanville, B.B., C.T. Best and R. Levenson (1977): "A cardiac measure of asymmetries in infant auditory perception", Developmental Psychology 13, 54-59.

Kimura, D. (1976): "The neural basis of language qua gesture", in H. Whitaker and H.A. Whitaker (eds.) Studies in Neurolinguistics (vol. 3), New York: Academic Press.

Kimura, D. and Y. Archibald (1974): "Motor functions of the left hemisphere", Brain 97, 337-350.

Kinsbourne, M. and R.E. Hicks (1979): "Mapping cerebral functional space: competition and collaboration in human performance", in M. Kinsbourne (ed.) Asymmetrical function of the brain, 267-273, New York: Cambridge University Press.

Klima, E.S. and U. Bellugi (1979): The Signs of Language, Cambridge, Mass.: Harvard University Press.

Lassen, N.A., D.H. Ingvar and E. Skinhøj (1978): "Brain function and blood flow", Scientific American 239, 62-71.

Levy, J. (1974): "Psychobiological implications of bilateral asymmetry", in S.J. Dimond and J.G. Beaumont (eds.) Hemisphere function in the human brain, London: Elek.

Martin, J.G. (1972): "Rhythmic (hierarchical) versus serial structure in speech and other behavior", Psychological Review 79, 487-509.

Mateer, C. and D. Kimura (1977): "Impairment of non-verbal oral movements in aphasia", Brain and Language 4, 262-276.

Molfese, D.L., R.B. Freeman and D.S. Palermo (1975): "The ontogeny of brain lateralization for speech and nonspeech stimuli", Brain and Language 2, 356-368.

Nakatani, L.H. and K.D. Dukes (1977): "Locus of segmental cues for word juncture", JASA 62, 714-719.

Ojemann, G. and C. Mateer (1979): "Human language cortex: localization of memory, syntax and sequential motor-phoneme identification systems", Science 205, 1401-1403.

Semmes, J. (1968): "Hemispheric specialization: A possible clue to mechanism", Neuropsychologia 6, 11-26.

Stevens, K.N. and S. Blumstein (1978): "Invariant cues to place of articulation", JASA 64, 1358-1368.

Studdert-Kennedy, M. (1975): "Two questions", Brain and Language 2, 123-130.

Studdert-Kennedy, M. and D.P. Shankweiler (1970): "Hemispheric specialization for speech perception", JASA 48, 579-594.

Studdert-Kennedy concluded by quoting Ludmilla Chistovich who as a conclusion of her report writes "We (our group) believe that the only way to describe human perception is to describe not the perception itself but the artificial speech understanding system which is most compatible with the experimental data obtained in speech perception research". He found that this was a very good statement of a heuristic programme, but emphasized that what is required is a constant interplay between the psycho-biological facts of the human behaviour and whatever robotic facsimile the engineers have managed to construct.

Hiroya Fujisaki summarized his report, giving a more detailed account of the first section on categorical perception based on slides illustrating his well-known dual coding model of discrimination. The fact that categorical perception appears in an apparent enhancement of discriminability on the phoneme boundary, and not in a suppression of discriminability within the category, was illustrated by reference to experiments with an r-l continuum presented to American and Japanese listeners. Categorization immediately after the auditory mapping and dominance of categorical perception on comparative judgement seems to be characteristic of the speech mode, but is also found in some cases of non-speech stimuli. Due regard should be paid to disturbances by noise (uncertainty) both in the categorical judgement process and in the retrieval process from the short term memory of timbre. The ability of categorical judgement is based partly on basic physical discreteness, partly on language specific criteria acquired through training in a specific language.

As for the perception of speech in context, Fujisaki emphasized that the importance of context can not be evaluated until we have studied the variability of phonemes in isolation.

Ludmilla Chistovich had been prevented from participating in the congress.

DISCUSSION

The discussion was opened by Kenneth Stevens, Sieb Nooteboom and Christopher Darwin.

Kenneth N. Stevens confined his remarks principally to the question of invariance versus non-invariance. It is obvious that when one produces phonetic segments in context, the articulators

have to move from one target to the next, and so the signal is clearly context-dependent. But if you examine the sound in the right way and look at the right places in the sound, you will see much less variability and more invariance for a given distinctive feature both in the context of other features in the same segment and in the context of adjacent segments. Stevens showed slides of the acoustic waveforms of the syllables ba,da ga,pa ta,ka. The samples were taken at the onset of the consonants and the spectra had been calculated in a specific way with a specific time window. He pointed out that in labials the gross shape of the spectrum was flat or falling and spread out in frequency. For the alveolars the spectrum was also spread out in frequency, but rising, or acute, and in velars it had a prominent peak in the mid frequency range. One may say there is compactness to the spectrum.

It is possible to devise algorithms or templates that will recognize each of these gross spectrum shapes - and the point is that if one looks at the gross spectrum shapes rather than at the details of where individual peaks are in the spectrum, one does see a considerable amount of invariance. Now, this is a physically measured spectrum with a linear time scale and with fixed bandwidths. What one should really do is to look at a spectrum as it is processed by the auditory system with the appropriate bandwidths and time constants of that system. At some level in the auditory representation that spectrum may well be influenced by what immediately precedes the spectrum. There are already neurophysiological data that would indicate that. The spectra would have to be brought more in line with what we know about psychophysics and the electro-physiology of the auditory system. But even at this acoustical level we see a measure of invariance for stop consonants, as far as place of articulation is concerned.

In this connection Stevens added some remarks on categorical perception. As one moves along the continuum from ka to ta, the auditory system does not treat the physical continuum as though you were moving continuously. As long as the sound has some sort of compact spectral peak it would sound pretty much the same, and it is only when this peak disappears that you will get a sudden change over to a different kind of sound. Stevens would argue that at some level of the auditory system there is some kind of unique response to each of the spectrum types characterized by the gross properties mentioned above.

Where should one look in the signal to find this invariance? Ludmilla Chistovich and Stevens agree that the places in the sound where there are rapid changes are the places which seem to contain a lot of information. If one looks at these places, one sees invariance not only for place of articulation but also for other distinctive features. The formant transitions are acoustic material that links these rapidly changing events with the relatively slowly changing events during the vowel. There is a tendency for a given phonetic feature to have invariant properties. Stevens would argue that the infant comes into the world endowed with mechanisms that are sensitive to these properties. It has a mechanism for classifying sounds, in particular features, as being similar. These relatively invariant primary acoustic properties help to define distinctive features and provide the signal with the kind of properties that enable the infant to learn language. The context-dependent effects which can go along with these primary properties can be used when necessary, perhaps in noisy situations or in rapid speech to supplement the primary cues.

Sieb Nooteboom had no disagreement with the description given by the reporters of the state of the art in speech perception research, but some comments with respect to the state of the art itself.

The underlying or most basic common goal of speech perception research is undoubtedly to understand the structures and processes by which a listener can recover from the acoustic signal what a speaker is saying to him. It is only when we have reached a basic understanding of speech perception in this sense that we can apply the insights gained to phonological explanation, improvement of synthesis by rule, etc. The most important of the processes involved may be labeled recognition. But experimental paradigms in our discipline draw heavily on forced-choice identification, discrimination, similarity judgements, and scaling, none of which studies recognition as a process in itself. In a typical recognition task each stimulus is presented once only and is potentially compared by the subjects with, for example, all possible words or morphemes in the language, whereas in identification stimuli are typically presented more than once and the response set is restricted by the task. With a very few notable exceptions (cf. Goldstein 1977, Marslen-Wilson and Welsh 1978, Cole and Jakimik

1978) recognition is not studied at all. In this respect research on reading, where considerable attention is paid to visual word recognition, is ahead of research on speech perception (Bouwhuis 1979).

Too much attention is focussed on phonemes and phonemic features at the expense of more comprehensive structures, words, morphemes, and prosodic structures, and their communicative function. For a listener to understand what a speaker is saying to him, he must generally recognize meaningful units. Words and morphemes are certainly the most important structures in speech perception. Most investigators seem to believe that once we understand how phonemes are extracted from the signal we can easily explain further linguistic processing. This is hardly true. We do not know whether word recognition is mediated by phoneme extraction, or rather, as recently suggested by Dennis Klatt (1979), by spectral templates, and we will never know until we turn to the study of word recognition. And even if word recognition turns out to be mediated by phoneme extraction, that is certainly not all there is to it (cf. the word completion effect in visual word recognition, Reicher 1969, Bouwhuis 1969).

The even more comprehensive suprasegmental or prosodic structures also contribute in several ways to a listener's recovery of what the speaker wants to say to him. It is a good thing that in recent years researchers have been paying more attention to prosodic structures. Attention has mainly centered around the connection between prosody and syntax, but Nooteboom thinks that two other functions are at least as important in daily speech communication. One is that differences in global pitch level, as well as the presence of normal intonational patterning, appear to increase the intelligibility of speech masked by speech (Brokx 1979). The other, and perhaps most important communicative function of prosody is to signal semantic focus (O'Shaughnessy 1978).

We should acknowledge that phonetics, and especially perceptual phonetics, has reached a stage in which it should not be limited to the study of consonants and vowels. Much is to be gained from widening the scope of the mainstream of our discipline.

References

Bouwhuis, D.G. (1976): Visual Recognition of Words. Unpublished Doctor's Thesis, Catholic University of Nijmegen

Brokx, J.P.L. (1979): Waargenomen Continuïteit in Spraak: het Belang van Toonhoogte. Unpublished Doctor's Thesis, Eindhoven University of Technology

Cole, R.A. and J. Jakimik (1978): "Understanding speech: how words are heard", in G. Underwood (ed.) Strategies of Information Processing, Academic Press

Goldstein, L. (1979): "Perceptual salience of stressed syllables", Chapter II of an Unpublished Doctor's Thesis, University of California Los Angeles, Department of Linguistics

Klatt, D.H. (1979): "Speech perception: a model of acoustic-phonetic analysis and lexical access", Journal of Phonetics 7, 279-312

Marslen-Wilson, W.D. and A. Welsh (1978): "Processing interactions and lexical access during word recognition in continuous speech", Cognitive Psychology 10, 29-63

O'Shaughnessy, D. (1976): Modeling Fundamental Frequency, and its Relationship to Syntax, Semantics, and Phonetics, Unpublished Doctor's Thesis, M.I.T., Cambridge, Massachusetts

Reicher, G.M. (1969): "Perceptual recognition as a function of meaningfulness of stimulus material", Journal of Experimental Psychology 81, 275-280.

Christopher Darwin started by quoting Ludmilla Chistovich (the same passage that is quoted by Michael Studdert-Kennedy at the end of his report). He concentrated his contribution on a discussion of the relation between computer speech recognition work and the human speech perception in the area of auditory feature extraction and phonetic segment identification.

The engineer does not have to make his system in a psychologically plausible fashion to make it work, but there does seem to be general agreement that speech recognition systems should take account of such relatively peripheral auditory phenomena as critical bands, middle-ear transfer function, growth of loudness and non-simultaneous masking although often the application to speech sounds has to be made on trust rather than on adequate psycho-acoustic data.

Chistovich, rightly, identifies as important the problem of how to represent the input parameters to an acoustic phonetic stage. She points out that theories of phonetic perception are going to be heavily influenced by the materials they have to work with. Thus, if speech understanding programmes are to be serious models of human perception we have to find ways of representing the input signal which are more psychologically plausible and more phonetically germane than a series of categorical labels representing the best, fitting one of a small (100-300) number of static spectral templates.

We have rather little idea what the parameters of an auditory representation should be. Probably it should represent all discriminable differences in the speech signal (taking the most liberal view of "discriminable"), rejecting none of the information to which the listener may need to be sensitive (cf. the work on early visual processing by Marr 1976), but on the other hand the representation must be explicit, organised along those dimensions that are most useful to subsequent processing. It is very different to state explicitly that, for example, there is a formant transition passing between two points in the frequency/time space than simply to represent the signal in a "neural spectrographic" form. The former requires extensive additional processing and important choices about what auditory dimensions to represent. These dimensions must allow not only phonetic classification but also the multitude of para- and non-linguistic decisions that we can make on a speech input, together with all those adjustments for speaker and rate of speech which bedevil recognition algorithms.

One property that a psychologically plausible auditory representation must have is to represent amplitude and spectral change explicitly rather than as a sequence of static events. Two experimental reasons can be given why this should be so:

First, the perceived loudness of a sound depends not only on its intensity but on the changes in intensity that precede and follow it. Jesteadt, Green and Wier (1978) have recently documented this effect which they call the Rawdon-Smith illusion after its co-discoverer (Rawdon-Smith and Grindley, 1935); they find that a rapid rise or fall in intensity is perceptually more salient than a slow change, so that subjects will, under suitable conditions match as equally loud two tones of the same duration and frequency that differ by 13 dB in intensity. Perceptually then, steady-states are (at least partly) defined by their edges, not vice-versa.

Second, the apparent perceptual spectrum of a sound is determined by the changes in spectrum that precede (and perhaps follow) it. Haggard and his colleagues (abstracted in Haggard et al., 1977/8) have shown that a flat spectrum can sound like, for example, [ɪ] if it alternates with a sound whose spectrum is the complement of [ɪ] (having zeroes where [ɪ] has poles).

As well as representing change explicitly, the auditory representation must allow auditory properties to be defined relative to a particular sound source. Silence, for example, is not absolute but rather a property of an assumed source. If a continuous formant pattern is perceptually divided into two assumed speakers by rapid alternations in pitch (Nooteboom, Brokx and de Rooij, 1976) then each speaker appears silent while the other is speaking and, with suitable choice of formant patterns, this perceptually induced silence can cue stop consonants (Darwin and Bethell-Fox, 1977).

Finally, Darwin wanted to make it clear that he finds the interaction between psychological theory and computer algorithm extremely stimulating. It is too easy for someone working with synthetic speech as a tool for investigating human perception to equate the auditory or phonetic dimensions used by the brain with the control parameters of his synthesizer. Trying to write an algorithm to detect, say, voice-onset time is a sobering experience for anyone used to generating beautiful synthetic continua. Algorithms applied to large quantities of natural speech are an invaluable complement to the necessarily restricted psychological experiment.

But if such joint perceptual and computer endeavours are to produce a theory of speech perception rather than a pot-pourri of micro-theories, each concerned with particular phonetic distinctions, we need to be more concerned with the general constraints on speech sounds. What is it that lets us hear as an additional extraneous noise the badly synthesized part of an utterance? Or what allows us to hear speech through a masking pattern that, on a spectrogram, deceives the eye (Liberman and Studdert-Kennedy, 1978)? The answer for some is in "directly perceiving" the articulation, but we are a long way from being able to write an algorithm that can directly perceive.

References

Darwin, C.J. and C.E. Bethell-Fox (1977): "Pitch continuity and speech source attribution", Journal of Experimental Psychology: Human Perception and Performance 3, 665-672

Jesteadt, W., D.M. Green and C.C. Wier (1978): "The Rawdon-Smith Illusion", Perception and Psychophysics 23, 244-250

Haggard, M.P., G. Yates, M. Roberts and Q. Summerfield (1977-8): "Onset and offset spectra in the analysis of complex sounds", Annual Report 1-2, M.R.C. Institute of Hearing Research, Nottingham, U.K., 12-13

Klatt, D.H. (1977): "Review of the ARPA speech understanding project", JASA 62, 1345-1366

Klatt, D.H. (1979): "Speech perception: a model of acoustic-phonetic analysis and lexical access", J. Phonetics 7

Liberman, A.M. and M.G. Studdert-Kennedy (1978): "Phonetic perception", in R. Held, H. Leibowitz and H.L. Tenber (eds.) Handbook of Sensory Physiology VIII, "Perception", Heidelberg Also in Haskins SR-50, 1977, 21-60

Marr, D. (1976): "Early processing of visual information", Phil. Trans. Roy. Soc. B. 275, 483-524

Nooteboom, S.G., J.P.L. Brokx and J.J. de Rooij (1976): Contributions of prosody to speech perception, IPO Annual Progress Report 11, 34-54

Rawdon-Smith, A.F. and G.C. Grindley (1935): "An illusion in the perception of loudness", British Journal of Psychology 26, 191-195.

Dennis Fry expressed his admiration for Michael Studdert-Kennedy's report and for the amount of ingenious experimental work covered by the report. He only wanted, as a supplement, to put forward what he considered to be some brute facts about speech perception seen from the point of view of the acquisition of speech. All reporters mentioned this as an important aspect, but only in passing.

The first fact is that the child always proceeds from the referent to the sound distinction, never the other way about. He is paying attention to something in his environment and that gives him the motive to notice a sound distinction. Therefore this use of acoustic factors probably depends very much on an attentional factor, perhaps more than on the capacities for making these distinctions (cf. Carney and his co-authors).

The second fact is that the child evolves his own acoustic cues. It is essential to remember that every individual is free to evolve his own cues. The only constraint is that they must lead him to the right decision, that is to say to be able to recognize the word or whatever it is that has come in.

This means that the child attempts to learn to deal with the phonetic or perceptual space which is engaging his attention, not the whole phonetic perceptual space, and he starts with very simple cues, expanding the system of cues, that is, developing a larger and larger part of the possible phonetic perceptual space as the different references and the distinction between them make it necessary to do so. - And this whole development goes through reception first. You have to be able to receive, to distinguish, before you begin to produce; there is interaction between reception and production.

Dennis Fry thinks that all this is learnt. The fact that in different languages you get very different modes of dealing with the acoustic input is crucial, and the fact that once you have learnt one language you have difficulty in perceiving distinctions not made in your mother tongue, also shows that these things are learnt. Fry is not convinced of the existence of invariants or of any substratum of universal stuff, perhaps with the exception of the ability to distinguish between silence and sound.

As for the interaction between perception and production we do not keep it sufficiently in mind that every human individual being is hearing a completely unique version of his own sounds. Therefore no human being can make a perfect, and not even a very good match between the sounds he is producing and what he hears from somebody else. It is therefore important that the child develops a cue system which enables him to deal with what comes in. When he sends stuff out, he has only to ensure through his feedback that he is implementing the cues which he is using to listen to somebody else. You have only this amount of match. - Therefore Fry rejected the idea of a motor theory, also in the form that listeners should have to infer something about the vocal tract of the other person. This is not necessary if the whole thing is done on the basis of these cues.

Björn Lindblom showed slides of a distance metric box and of a block diagram of auditory analysis inspired by Manfred Schroeder, which, starting from a harmonic spectrum, converting the frequency scale into a Bark scale, and adding an auditory filter and a masking pattern, leads to two quasi-auditory excitation patterns, a quasi masking pattern and a loudness-density pattern. In accordance with Plomp he thinks that the perceived difference between two static stimuli depends on the area between two curves in the auditory excitation pattern. On this basis he and his co-workers try to explain: (1) the F2' data that have come out of the experiments by Carlson, Granström and Fant (in this respect the results are very positive), (2) Flanagan's difference limen data (which Lennart Nord has had some success in explaining), (3) dynamic events, e.g.: Is a vowel formant target identified better in a

dynamic than in a static context? (Karin Holmgren has found that it is not.) This latter result is not totally in agreement with the point that Darwin made, i.e. that the human speech perception mechanism is primarily sensitive to changes, although Lindblom, generally, agrees completely with this point of view.

Lloyd H. Nakatani agreed that phonetic perception is fundamental to speech perception and that, as Studdert-Kennedy said: "Perhaps all these years of studying C-V syllables have not been wasted after all", but now it is important to concentrate more work on prosody and bring more linguistic facts in. In prosody the cues are complex, and there are great idiolectal differences between talkers. We cannot continue generalizing from the Haskins speech synthesizer to the whole population. In some recent papers in JASA a new technique that attempts to cope with more complex perceptual phenomena has been described.

Dennis Klatt emphasized that you should not set up a dichotomy between phonetic segmentation and the possibility of going directly to larger units, like the word. Both phenomena are well motivated. Phonetic segmentation is supported by the fact that the speech production process manipulates units such as segments, and by the fact that one must have a method for understanding new words. But going directly to the word restricts the phonetic strings to look for and helps solving ambiguities. It also helps to interpret durational cues, because, e.g., stress plays a role. One possibly has to build into our model of the perceptual system kinds of constraints that will make for optimal decisions.

Klatt's second point was that there is no logically necessary connection between feature extraction and phonetic labelling. The features may lead directly to words. One should investigate the feature problem by building very simple models of perception, trying if simple psycho-acoustic distance metrics can be used to make predictions of the sort that are made by phonetic data or not. If not, it points to feature detectors. Probably some of the natural quantal categories will come out of very simple assumptions about the peripheral system and the distance metrics.

The context effects mentioned by Darwin will be troublesome for distance metrics, but this does not prevent a solution. The distance metric is going to be a change-over-time kind of metric.

Osamu Fujimura mentioned a recent study at Bell Laboratories by Marian Macchi treating the role of consonantal transition in perceptual identification of vowels which has been published in Speech Communication Papers edited by J.J. Wolf and D.H. Klatt 1979. In contrast to what Strange et al. reported (JASA 60, 1976, p. 213-24), Macchi's result demonstrates that vowels in isolation can give rise to a very high accuracy of identification when appropriate care is exercised concerning dialectal problems and the possible difficulty in orthography (Macchi used rhyming tasks instead). It is possible that dialects vary considerably in the phonetic characteristics of gliding, even for so-called monophthongal vowels in English, and these gliding effects are particularly important in the case of isolated vowels as opposed to syllables ending in a consonant, because the VC transition in the latter case reduces or perceptually obscures such gliding effects.

Dominic W. Massaro: It is recommendable to utilize an information processing approach in speech perception, because the goal of this approach is to delineate the stages of processes that occur between the acoustic stimulus and the meaning in the mind of the observer. It has been found that even at an early stage of processing where you are taking raw feature information and integrating it together it is necessary to incorporate what the listener knows in terms of speech he or she has heard before, in terms of constraints in the language, and in terms of possible words or nonwords and so on. So even at this early stage we have to develop models that allow the contribution of higher order processes. Rather than opposing bottom-up and top-down processes, what has to be developed are specific formal models that describe the integration of both sorts of information.

As for features Massaro has found that they are not binary. In fact, listeners have knowledge about the degree to which a feature is present in the speech chain.

Pierre L. Divenyi took up the problem of categorical perception as treated by H. Fujisaki. He found that the problem whether perception, and categorical perception in particular, is articulatorily or auditorily bound is an artificial one. In Fujisaki's second stage there may even enter non-speech auditory events. At the higher stage of perception there is no time for a detailed analysis. Categorical perception is a result of applying an

a priori decision process about what to pick from the signal, and this results simply in discrimination peaks and categories.

Steve Marcus argued that intermediate levels between the acoustical signal and the perceived word are only hypothetical constructs. It appears from split-brain studies that in the right hemisphere word recognition is obtained by an acoustic-lexical mapping system. It would be parsimonious to assume that the left hemisphere used the same system, and that the further possibility of the left hemisphere for segmental analysis would be used for special tasks only, such as CV-recognition, rhyme detection and learning of new words. An intermediate stage seems to be necessitated by current work on the combination of acoustic and visual-articulatory cues (lip reading) in speech perception. It would be interesting to examine whether split brain patients can use lip reading.

Secondly, Marcus argued that there is no empirical justification for assuming a phonemic level. It could also be a continuous real time integration, perhaps using some temporal reference points, which may be purely acoustically determined. The fact that initial phoneme detection times are dependent on factors affecting word recognition speaks against the role of phonemes in perception.

Herbert Pilch. Like Sieb Nooteboom H. Pilch regretted that the study of speech perception has been limited to controlled responses to synthetic stimuli. Our goal must be to understand speech perception in routine communication.

Prosodics signal neither syntax nor sentence meanings, but discourse structuring in the rhetorical sense. Monotonous reading fails to achieve communication, whereas intact prosodic performance can outweigh severe aphasic disturbances in phonemes and syntax.

Routine perception works on the basis not of specific linguistic elements (such as phonemes, syllables, words, sentences) but of total messages. Minimal distinctions may be hard to grasp.

The listener may, however, shift the focus of his perception from the total message to any particular element, i.e. perceive the speech signals as (a) a message, as (b) a linguistic structure, or as (c) noise. In case (a) he may miss the message, in case (b) the structure (cf. H. Pilch: Auditory Phonetics, Word (in print)).

James Pickett: Taking up Studdert-Kennedy's hypothesis that we perceive the speech movements directly, Pickett proposed that we should attempt to set up features of movement (What is moving? where is it going? how is it moving? how is it related to preceding and following movements?) and see where it leads.

Adrian Fourcin: Referring to Dennis Fry's contribution Fourcin confirmed that children do indeed go from the recognition of very simple physical features to levels which are more recondite and varied in the spectral form of the signal. So with the voiced-voiceless opposition you go initially, in the earliest years, from three to five, from a skill of discrimination based on whether voicing is there or not, to a skill based on the onset of the first formant as flat or rising.

Children who are totally deaf can learn to produce clear stress contrasts by means of a visual display of auditorily relevant information. Moreover, by using an auditory pattern approach and giving them an electrical stimulation of the cochlea you can teach totally deaf children to make discriminations based on their pattern knowledge and give them a categorical ability to discriminate which is not at all based on any motor references.

But in order to communicate at a fast rate you have to use a sort of parallel processing technique which is necessarily dependent on your knowledge of coarticulatory constraints.

T.M. Nearey reported that Assmann (cf. vol. I, p. 221) obtained the same results as Marian Macchi (see Fujimura's contribution to the discussion), i.e. a much higher recognition of isolated vowels than should be predicted according to Strange et al., when factors of dialect, orthography, etc. were controlled.

Hiroya Fujisaki emphasized that the role of prosody may be quite language specific. Further, he showed a number of slides illustrating his acoustical and perceptual investigation of Japanese accent.

Michael Studdert-Kennedy concentrated his final remarks on four points:

1. The problem of recognizing dynamic vowels against isolated ones is very complicated. O. Fujimura has showed that centers of vowels extracted from running speech are not readily identified and do need the surrounding formant transitions. Percent correct identifications is probably not the most sensitive measure for that question.

2. Studdert-Kennedy had not attempted to argue that we have no acoustic property detectors. Presumably there is some system within the brain that is able to pick up acoustic properties, but the question is whether there is any grounds for supposing that those property detectors are opponent detecting systems, and whether there is any ground for supposing that they have been adapted for linguistic purposes. In this regard he would rather go with Kenneth Stevens and suppose that language is simply exploiting properties of the auditory system rather than the other way around.

3. In answer to Steve Marcus: To what extent you use auditory contours in listening is an open question. But Studdert-Kennedy would give most of Marcus's data an exactly opposite interpretation. For instance, the fact that phoneme recognition comes after word recognition has nothing to do with perceptual processes, it is a question of experimental tasks and of bringing things into consciousness.

4. Studdert-Kennedy found the data on child language acquisition very important, for instance the work by Boysson-Bardies and by Lise Menn. Another field of research which is highly relevant for the problem of speech perception is that of sign language. Many of the processes of acquisition resemble quite closely the processes of acquisition of spoken language which suggests that what we are dealing with is a very general system that is highly flexible and adaptable to a variety of different circumstances.

REPORT: PHONOLOGY
(see vol. I, p. 103-152)

Reporter: Hans Basbøll
Co-reporter: Stephen Anderson
Co-reporter: Joan Bybee (Hooper)

Chairpersons: William Haas and Kenneth L. Pike

REPORTERS' ADDITIONAL REMARKS

Hans Basbøll: On an abstract level of discussion, it is very hard to disagree with Anderson's claim that one should avoid a priori statements about psychological reality and other linguistic issues, as well as "the arbitrary imposition of restrictive principles which rule out otherwise well-motivated descriptions" (p. 142).[1] I also fully agree with the claim that formal questions just like other scientific questions should be taken seriously.

I am in agreement with the claim that the very fact that part of the traditional field of study cannot be dealt with adequately within a certain framework is not a decisive argument against the use of that framework in other parts of the field. Thus I would suggest that the SPE approach towards markedness, which is considered quite unsatisfactory by both of my fellow reporters, can in principle be used in a rather specific subpart of that subfield of the study of sound structure which it was devised to deal with: namely, to account formally for implicational universals à la Roman Jakobson between sound types. What is outside the scope of the SPE approach towards markedness and similar approaches are other aspects of natural systems and natural segments (like prohibited segments and contrasts, or internal economy) as well as explanation, in any interesting sense, of the relation between phonology and phonetic substance. Such an explanation remains an important task of our discipline, of course.

While I also partly agree that certain efforts of Natural Generative Phonology might be termed reductionist, namely the axiomatization of strong constraints on the form of grammars, I would, on the other hand, suggest that a considerable part of the

---

1) Pages refer to volume I.

efforts of SPE-phonologies is reductionist in the sense that large amounts of evidence, and thus potential counter evidence, is not taken systematically into consideration. The data considered as evidence is too often limited to a static set of occurring forms as against all the facts which the languages present (cf. p. 142), including those that may be revealed in psycholinguistic experiments, and in studies of language acquisition, language loss, and so on.

What is really at issue are two related fundamental problems: first, the question of predictability and second, the relation between model and reality - in particular: What is the model a model of? and how can it be tested?

I would like to emphasize that in my report I have not stated nor implied nor suggested that the goal of phonology is complete predictability (compare also Labov's variable rules which are probabilistic rather than deterministic). I have said, however, and that evidently is not very new, - that a scientific description should be prognostic in the sense that "it should make predictions (which in principle could be refuted) about something outside the material on the basis of which it was constructed in the first place" (p. 117). That phonology could or should in principle be deterministic is a claim which would hardly be defended by anyone to-day, with the possible exception of a few radical behaviorists. I also think that most linguists would accept the hermeneutic goal of "ex post facto understanding" (p. 140), at least faute de mieux. I certainly also agree that the identification of mutually inconsistent principles may advance our knowledge (for instance the "internal" vs. "external" economy of sound systems according to Martinet), but in such cases our efforts should be directed towards finding constraints on the principles in question to diminish (or better, remove) the field of conflict between them. That a phonological description or theory should be prognostic, on the other hand, is a necessary condition for its being even partly tested for falsifiability, that is for one type of decision on how it relates to "reality".

What the model or theory is a model or theory of is, of course, a vexed question which is closely related to the issue of the reality of phonological descriptions in general, either psychological or sociological. I shall not go into that matter here,

but only briefly remark first that the frequently used phrase 'linguistically significant generalization' may have very different meanings according to the type of reality - if any - ascribed to phonological or other linguistic descriptions; and second, the question of psychological reality is not of the yes-no-type, but there would be a whole scale of possible relations between some internal grammar and an observationally successful model of it (as far as its output is concerned), stretching from a "black box" to a point-to-point-correspondence.

The relation between model and "reality" is of a dialectic nature: The model specifies a number of theoretical constructs, like "natural class" in the "model-internal" sense, defined as a certain set of co-occurring distinctive features, to take just one example. At the same time, real languages present natural classes of segments in the "model-external" sense, that is sets of segments that function as a class in real processes in languages, be it acquisitional, synchronic, diachronic, or whatever. The testing and modification of this part of the model is then a series (generally an infinite one) of steps whereby the sets of segments specified by the "model-internal" and "model-external" natural classes should be brought to coincide, while still respecting all other conditions on the theoretical constructs, such as other types of criteria for the establishment of distinctive features. The model specifies which types of data we should look for, and also which aspects of the data should be considered pertinent and which aspects irrelevant; it must then be independently decided whether the data is in conflict with the model or not.

Now, the point is that this partial testing procedure presupposes that the parts of the model not under consideration for the given purpose must be treated as given for that purpose (as I have said in my report): you cannot test everything at the same time. This is all right if the scientific paradigm within which you work is accepted as basically correct in its main lines, and that is exactly where a clear and fatal division of attitude towards the state of the art occurs, in particular whether the "conceptual richness" of SPE in Anderson's words (p. 136) corresponds to anything outside the model itself. Some people, like my fellow reporter Stephen Anderson, think that SPE represents

"monumental results" (p. 138) and that it is methodologically sound whereas others, including myself, consider SPE - despite its monumental efforts and certain merits - as misguided in quite fundamental respects. I should like to stress once more that both of these two attitudes towards a research paradigm may per se be scientific.

Stephen Anderson: I will focus my attention on the apparent conflict between rationalist and empiricist approaches to sound structure, this being a distinction that I think is at least operationally similar to that raised by Basbøll as the distinction between formal and substance based approaches. This distinction can usefully be approached in terms of the following question: After we have taken into account all those aspects of speech that are associated with more general problems, and which can be approached from outside the domain of language per se, how much is left? Substance based views have typically pursued the possibility that virtually all aspects of language are accessible from one or another more general point of view, and that they can be treated as special cases of the functioning of the articulatory apparatus, of generalized perceptual strategies, of general limitations on memory and processing, and the like. As a result, these researchers have put a great deal of faith and emphasis on the possibility of experimental verification of the details of linguistic structure, for example on the devising of psychological tests to determine on the basis of constructed tasks whether particular proposed phonological rules are psychologically real or not. The substance based linguist takes the absence of such external evidence as establishing a case ex silentio against the proposed analysis as a correct account of language.

The formal approach, on the other hand, has been motivated by the feeling that there are distinct aspects of language which are proper to itself, not studyable necessarily as special cases of other systems. Hence, for the formalists, the absence of direct external accounts for some area of language is not very surprising, or a cause for alarm. This is because this line of reasoning allows specifically for the possibility that among the interacting domains that contribute to the facts of speech, we may find a language faculty which is not indeed reducible to features of other kinds. If so, there is no reason, in principle,

to expect that such a language faculty, if it exists, ought to be directly accessible to inspection in other terms, through constructed psychological experiments of a given kind, for example. The validation of claims of this sort then, would rest not on the establishment of direct links between them and external observables but rather on the inferences that can be drawn from the success, or lack of it, which they achieve in facilitating and revealing regular connections among phenomena, and in uncovering orderliness and coherence within the complexities of languages.

It is important to see that the primary issue between these two views, that of the existence of a specifically linguistic aspect of cognitive structure, not accessible in other terms, could probably never be settled conclusively. One might, of course, establish that a given aspect of linguistic structure is a special case within some more general demand. However, if we construe the proposal that there are aspects of language which are systematically not studyable in such terms, we construe that proposal as an empirical proposition about the nature of language. It is hard to see such a position as other than completely mystical in the extreme. This is, however, not really a matter of empirical fact, but rather a matter of choice of research strategies. Whether or not one ought to limit the terms of linguistic description to elements that can be given an external foundation. As a matter of choosing between research programmes, it seems to me that the claim that all aspects of linguistic structure ought to have some more general basis and ought to be accessible from some other realm, is at least equally mystical, at least in the absence of any such account from any area of linguistic phenomena. The best way to motivate the decision on this issue is to attempt to establish not the correctness but the plausibility of one or the other position. One does this, of course, by demonstrating the ability of this position to provide satisfying and detailed accounts of regularities among the facts of natural languages.

To my mind, the formalist, or as I would prefer to say, the rationalist approach has much the better track record in this regard, though I am sure there are many who will disagree with that. Nonetheless, I hope to have suggested that the choice is by no means an obvious one and in particular, that the formalist pro-

gramme is in no way vitiated, as is sometimes suggested, by its indirect relation to surface facts; that is indeed its essence and its greatest interest.

Joan Bybee Hooper: In the transformational generative tradition a working hypothesis seems to be that if X and Y show some characteristics in common, then they must have the same underlying form, so this produces an emphasis on similarities among elements and has led to a dismissal, occasionally, of surface differences. The results are hypotheses that are untestable because it is always possible to invoke what Botha calls blocking devices, caveats that put hypotheses beyond the surface phonetic facts. This position is exemplified by SPE. The contrary position, which is the one that I accept, requires that linguistic hypotheses be testable (either by comparing them with the surface forms of language or by some kind of experimentation). This is not an a priori constraint on a theory of phonology, it is a different way of approaching facts. Nor is it an attempt to do phonology without an appeal to any abstract entities, because, in fact, all phonology is abstract.

Basbøll expresses the opinion that there is not a big division among these two approaches to phonology. He says in his written report that they share common bases of argumentation and understand each other reasonably well. It seems to me that this is not always the case. There is not a single set of shared assumptions and, in fact, some misunderstanding does ensue. In his paper, Stephen Anderson presents an example from Javanese, intended to falsify the claim that morpholexical rules should apply prior to purely phonological rules. But all we can conclude from the data is that the morphological rule must apply to basic adjectives with round vowels in final position. Only if we assume that lexical representations cannot contain any information that is the output of productive rules does it follow that the morphological rule must apply after the phonological rule. If we do not make such an assumption, the example shows that lexical representations, i.e. the phonological representations relevant for word formation, contain predictable phonetic detail, or to put it another way: the lexical representation has been restructured to contain the output of productive phonetically conditioned processes. The example shows an important difference between the

two approaches: in generative phonology it is assumed that underlying representations are negatively defined by the rules, but I believe that underlying forms and rules can and should be determined independently of one another by examining various types of linguistic evidence and independent or non-structural evidence.

In a paper by Donegan and Stampe in the volume edited by Dinnsen from the Bloomington phonology conference, they characterize a theory of natural phonology by saying: "This is a natural theory in the sense established by Plato in the Cratylos, in that it presents language as a natural reflection of the needs, capacities and world of its users, rather than a merely conventional institution. It is a natural theory also in the sense that it is intended to explain its subject matter, to show that it follows naturally from the nature of things. It is not a conventional theory in the sense of the positivist scientific philosophy which has dominated modern linguistics in that it is not intended to describe its subject matter exhaustively and exclusively, i.e. to generate the set of phonologically possible languages." This characterization has two parts: The first one deals with the difference between whether the explanation for linguistic structure will come from general properties of human users of language, or whether it is contained in something that is specifically linguistic and not accessible to verification (although it is not clear to me how this specifically and uniquely linguistic thing is immune to experimental investigation). Secondly, they say that the goal of a natural theory is not to produce exhaustive descriptions of its subject matter. It seems to me that trying to meet the goals of observational and descriptive adequacy has often forced us into making unwarranted theoretical decisions which we may at the time characterize as arbitrary, but in fact then we accept them and never go back to reexamine them; however, such assumptions should be reexamined in view of empirical evidence. Notation is the tool of a theorist and should not be mistaken for the theory itself.

DISCUSSION

Charles-James N. Bailey, Edmund Gussmann, and Henning Andersen opened the discussion.

Charles-James N. Bailey: Basbøll stresses the role of prediction and explanation. But he does not observe that development is what explains states and their structures; states cannot predict anything but what is in their own scope, and they can explain very little. For minilectal linguists - those who posit idiolects as the object of linguistic investigation and accordingly limit their models to static models - logic suggests that they should give up the goal of exact prediction.

Stephen Anderson's position is quite consistent with his synchronic orientation. He claims that markedness is getting vaguer; but developmental linguistics has been able to define naturalness and markedness quite exactly. Two kinds of dynamic data are relevant for defining the natural and for analysis and description: dynamic changes and comparative patterns (pattern is created by the dynamic principle). With the anticomparative models of minilectal linguistics - phonemes, idiolects, dialects, etc. - the theoretically interesting aspects of linguistics are virtually ruled out, for they demand comparative analysis: naturalness, child language, historical and dialectological linguistics, etc., which are all excluded on principle according to the definitions of phonemes, idiolects, etc. To study development with static tools would be worse than trying to drive a nail with a screwdriver. Since patterns of development are gradient, non-gradient tools are likewise fairly useless. One cannot even describe the morphology of German nasal-stem masculine nouns adequately, for example, with non-gradient models.

Aside from gradience, larger conceptual differences separate the underlying segments of three theories: (1) The classical (taxonomic) phoneme was neither internal-reconstructive nor comparative. (2) The generative phoneme is internal-reconstructive but not comparative. (3) The phoneteme is both internal-reconstructive and comparative, or polylectal. Only the latter is valid for development (comparative tasks, including child language acquisition), for theory, and for pedagogy. Development has two sides. One is the inner-linguistic side, where explanations in phonetology (dynamic phonology) must be sought in phonetics and ultimately in anatomy and bioneurolinguistics. The other side is the social side: a development must not only come into existence among children, but must also be adopted by others if it is to survive. Developments due to social or extralinguistic causes may be natural-like, or they may be, and often are, unnatural as in the borrowing of older or of foreign forms, hypercorrect rule-inhibitions, etc. This side of language is only semi-theoretical since many of the relevant conditions are hardly predictable, though creolistics is getting better at predicting changes under different social conditions and with different types of linguistic mixtures. Since Stephen Anderson seems to have a rather negative view toward extralinguistic explanations as well as doubts about some of the explanatory achievements of phonetics, he seems to be skating awfully close to advocating an YROEHT instead of a THEORY; An YROEHT predicteth not; - neither can it explain.

Since it is clear that some linguistic developments are natural and that some are not, and since all languages are mixed and have both of these elements, the immediate goal of linguistics ought to focus on understanding only natural developments and leave the rest for the future.

The abstractness controversy is merely an off-shoot of the really fundamental issue, namely, what are the facts to be analyzed? Our differing views on what is really real affect our views on what data are really relevant to linguistics. If I say that languages have both natural and non-natural phenomena, and you disagree, how could we ever agree on what data are to be admitted or excluded from linguistic analysis?

Even in connection with derivative matters there are several issues of phonetological analysis which are more fundamental than abstractness: There are reasons for believing that instructions from the central nervous system to the articulators are bundled differently in syllable-timed languages and in stress-timed ones, viz. in syllable-sized units and in measures, respectively. One of the deepest issues today is to specify the differences between phonomorphological and morphophonic (phonetological) rules. Another matter of interest is the fact that the segmental and suprasegmental uses of prosodic features are different: several

rules of English are respectively forwarded and hindered by these different functions of length.

Stephen Anderson takes the wrong view towards different historical developments and their use in the erection of a predictive theory. The difficulties exist only if one excludes the appropriate answer and mechanism: creolizing substrates and superstrates.

If you deal with idiolects, you can always say: "that is your idiolect, not mine", which effectively excludes both proof and replication - and theory. The best way to do linguistics is the way children and adults "do languages", viz. polylectally. Theory - if it means explanation and prediction - depends on development and change, on ascertaining how structures come into being, and on a dynamic comparison of the variation patterns resulting from change. We must admit that it is development that explains states, not vice versa, and then either give up all hope of synchronic explanatory theories, or become developmentalists. This is the paradigmatic difference among frameworks today.

Edmund Gussmann: The so-called substance based approach is in fact also a formal approach, but formal in a different sense. In natural generative phonology certain theoretical restrictions and conditions are established on the basis of some external evidence. But then these restrictions are generalized and applied to other data for which no external evidence is offered or simply where the evidence is not available. This is, of course, perfectly legitimate, but it shows that Basbøll is not right in what he says in footnote 8 of his report. In fact, substance based phonologists proceed in exactly the same way as abstract phonologists, though their restrictions are largely phonetic. But this phonetic nature is, in fact, often avoided without any real justification. For example, the "true generalization condition" is exempt from applying in the case of different styles and tempos.

When professor Hooper claims that phonological rules should correspond to phonetic data in a predetermined way, then there is little for descriptive or practising phonologists to do, since we have here really some sort of discovery procedure.

The standard generative approach to the question of how much structure should be assigned to individual lexical items was autonomous by being divorced from rules of word formation. A number of problems could have been avoided, if the direction of morphological processes had been taken into account. In some instances you can show that the rules of word formation have to take as their input the surface phonetic representation, in other cases the data argue just as strongly for abstract underlying representations as their input. There is a general non-existence of a theory of word formation. Here English seems to be a bad language to start with. In Slavic the very common expressive formations, such as augmentatives, diminutives, which are highly productive, are morphological processes which involve a number of phonological consequences. These should be studied in the first place, and rather than wondering whether 'serene' and 'serenity' are related. It is precisely in the interface of morphology, both inflectional and derivational, and phonology, that one should seek justification of phonological generalizations rather than in arbitrarily imposed restrictions of any sort.

Henning Andersen: Stephen Anderson's report seemed to me a very gracious concession of the total defeat of TG phonology. His remarks today seemed to contrive admission that it has not produced any results as a consequence of the monumental efforts made.

Basbøll's choice of leaving aside the vast amount of papers and monographs that contain important theoretical contributions under language-particular headings is regrettable. As to his limitation to descriptive linguistics, Bailey has taken care of that. But when Basbøll, in one of his footnotes, defines the substance based approaches as ones that go beyond the normal use of language, he must mean by that that they are interested in real data, meaning the use of phonology in speech, including speech errors, in verbal games, in poetics, by children, by aphasics, and so on.

In the same footnote, 'substance based' does not mean 'substance based' but rather 'speech based', - the traditional distinctions between language and speech, form and substance, etc. should be maintained also in discussions of these issues. I would like to ask Basbøll and Hooper to clarify what they mean by the distinction between formal and substantive, or if they understand them as being as vague as I do.

It is important to understand that language is something which is constantly changing, whose existence is in transmission from speaker to speaker, from generation to generation. Synchronic analysis is an artefact of the analyst. One must not identify

synchrony with the static, nor dynamism with diachrony: there can be dynamism in synchrony, and in diachrony you can talk about static facts, viz. the correspondences between two stages of a language.

In the transmission of language there are two logically distinct processes at work: deduction and abduction. Speakers know the grammar of the language and can produce deductively utterances which are correct. If you know the grammar, you can predict what sorts of utterances are going to be produced by that grammar. The other phase is the abductive one, by which speakers (children or adults) infer the grammar of the language from the speech they hear from speakers of the same dialect or from other dialects or even a foreign language. Logically, this is a process of hypothesis-making, about the content of the speech or about the grammar behind the speech. In this phase we cannot predict, but we can somehow understand the grammar. You cannot predict a grammar from the data, but you can form hypotheses about it. When we have constructed a grammar and understand that as a hypothesis, we can predict what sorts of innovation will be acceptable to speakers of that language, what sorts of verbal games will have which results, what kind of specific data would arise in aphasia - and we can test these hypotheses. On the other hand, given the speech data that learners of a language face when they acquire the language, we cannot predict the shape of the grammar they will produce. But we may be able to approach something like prediction if we understand that what they have to do in the process of arriving at a grammar is to make decisions, to form hypotheses. And if we understand that the data is susceptible to diverse analyses, contains ambiguities, we can capture these difficulties of analysis by formulating alternative hypotheses, and these hypotheses can then be subjected to empirical tests.

A proper theory of the ontology of language, which will be a proper theory both of synchrony and of diachrony, will enable us to both predict and to understand, will enable us to explain in both the senses that Bailey used, and hopefully future contributions of this kind will take in a wider scope of the field and see to what extent these various issues are faced by people working not specifically on descriptive linguistics but also on historical and pathological aspects of language, as well as the contributions made by people working in language-particular fields.

Joan Bybee Hooper: Gussmann says that if rules correspond to the phonetic substance in a predetermined way, then there is nothing for phonological theory to do. I think that is wrong. The formal theory may tell me what a rule is, given the phonetic data, it does not tell me how to figure out why there are these rules in particular rather than the other logically possible rules.

A clarification of the notion of substance: As an example we could consider the kind of criteria used in phonemic description; there are distributional criteria and then there is the criterion of phonetic similarity. Phonetic similarity would be a substantive criterion, while distribution would be considered formal. Another example: morphophonemics based on the properties of a morphological system would be a substantive approach, while morphophonemics treated as phonological would be a more formal approach.

Hans Basbøll: Synchronic linguistics seen as something absolutely static is a conception which I would not share.

Stephen Anderson: My view of the state of the SPE programme is that it proposed a particularly ambitious goal for constructing a logistic system that would reconstruct all of the content of sound structure. Certain fundamental inadequacies were clearly revealed in the comprehensiveness of the goals of that programme, as phonetic substance came to be taken more seriously into account. It seems to me that reactions to the perception of these failures have tended to throw out the baby with the bathwater and abandon the entire programme of SPE, and in particular its underlying rationalist assumptions, in an attempt to provide a rather radical sort of therapy for these problems. It seems to me that that is an overreaction; that one does indeed want to recognize that there are inadequacies in the attempt to reconstruct in such a logistic system all the content of phonology, but, nonetheless, one wants to preserve for that sort of system a central role in the development of phonology much as the sort of system in the Principia serves as a fundamental object of study within metamathematics.

Victoria Fromkin: The question is not: is the theory formal or substantive? but rather: is it a true theory of human language? I think that what Stephen Anderson has been trying to say is not that questions of articulation, etc., are not necessary for understanding certain aspects of language use, but that it is not necessarily the case that all aspects of language can be accounted

for by reference to these other aspects of language production and perception, etc. These questions of the philosophy of science are important because they have led us to look at different aspects which, hopefully, will eventually lead us to understand the nature of human language.

John J. Ohala: The issue of the psychological reality of phonological constructs has been raised during the discussion of this report and, in my opinion, has been made unnecessarily complex. I would like to simplify it with the following analogy, which is designed to appeal to the many academics in the audience. The problem of assessing the psychological reality of phonological constructs is very much like the problem the teacher faces in trying to verify that a student has mastered or knows the subject matter he has been exposed to in classes. How can this be done? Let us imagine three approaches: the teacher that takes the 'formalist' approach will just speculate on what it is possible for a student to know and will assume that that is what all students know. The teacher who would have most in common with those phonologists who have here been characterized as accepting 'substantive' evidence, would rely on additional 'external' evidence of a student's knowledge, e.g., what books he had in his library, whether he nodded sagely during the teacher's lectures, laughed at his jokes, etc. The teacher who would take the experimental approach would demand of all students some behavioral evidence that they had mastered the subject matter, e.g., performance on a written or oral test, an original paper or thesis, etc. Naturally this performance should not be attributable to anything other than the student's full mastery of the subject, e.g., cheating or random selections of answers to 'true/false' questions. I leave it to all those academics in the audience to decide which approach they would use. I would hope that whatever decision they make, however, that this would influence their practice in phonology, too.

The point is that different types of evidence in phonology vary considerably in their ability to unambiguously tell us what is in the speaker's head. Most of the evidence characterized as 'substantive' in this discussion, e.g., speech errors, sound change, is quite ambiguous in this regard. Only evidence from tests (experiments) can be minimally ambiguous. This is not to say that there cannot be a bad test. But the proper response to a bad test - both in academia and in phonology - is an improved test. Teachers expend considerable time and imaginative effort refining the tests they use to assess the psychological reality of students' knowledge. Why shouldn't similar effort bear fruit in phonology?

Natalie Waterson: I should like to draw attention to another theoretical approach: to Prosodic Phonology initiated by J.R. Firth in England. Very briefly: most phonological theories have phonemic segments as the basic units of description, whether explicit or implicit, yet there is general recognition by those who study speech perception that the phoneme has yielded little in the way of insights to our understanding of how speech is perceived and interpreted, and it is becoming plain that it is not the right unit for such studies. In Prosodic Phonology the unit of description is the word, phrase, or sentence, and features which synthesize the word, etc., into a whole as well as those that divide it up are taken into account, i.e. syntagmatic and paradigmatic relations.

The phonological system of a language is thus described in terms of different word, etc., structures and not in terms of a system of phonemic segments. No exposition of the theory is available but there is plenty of illustrative material in theses and papers produced in the Dept. of Phonetics and Linguistics, at SOAS, University of London. Most of the material is about Oriental and African languages and the only English material are my papers on child phonology.

It is interesting to see the influence of Prosodic Phonology on developing theories, for instance on Joan Bybee Hooper's approach, and autosegmental phonology.

Richard Coates: The SPE type of phonology, represented here by professor Anderson, has tended to specify a kind of codified norm, whereas professor Hooper's system specifies the linguistic rules which would characterize usage as being the starting point of changes. I think that together they comprise the native speaker's system, both a kernel, or norm, available to him, and a system of partly specified potential directions of the changes. Thus, the output of morphology would not be absolutely rigidly defined, and we may imagine a speaker who makes very few morpho-

logical connections between surface forms not connected by phonological rules, on the one hand, and on the other a speaker who fluently manipulates a morphological and phonological system (à la James Foley's native speaker).

Wiktor Jassem: Fifteen years ago, or more, three points were made about generative phonology: observational adequacy, descriptive adequacy, and explanatory adequacy. Now, in the old days so little observation was done that it is difficult to say whether it was adequate or not; descriptive adequacy described rather what was going on in the minds of the theorists; explanatory adequacy, for which the criterion was simplicity, led to rules which in structural phonology could be expressed by three or four symbols but which in TG took complete pages so full of things that you could not see the wood for the trees. My point is: I suppose that revolution in phonology did not start twenty or seventeen years ago with Chomsky, - revolution in phonology, according to what I have heard today and read in the Proceedings, is starting now!

Royal Skousen: Each approach to phonology proposes a method of analysis. In some sense they are all formal in that they look at the data and attempt to derive a description from the data, but I would prefer to call that a method of induction or learning. I would like to suggest that, in addition to these formal considerations or these principles of learning, there is a need also for an empirical interpretation of the description: What does my description actually predict about language usage, about language intuition? - Furthermore, we need first to explicitly determine how we get our description from the data, and secondly, to answer the question of what would convince us that our description is right or wrong, because in the absence of such arguments we do not really have a theory at all.

William Haas: There is another kind of opposition that has to be reconciled, namely the opposition between empirical and speculative. More than twenty years ago, Martinet published his "Phonology as functional phonetics". And that was a kind of reconciliation: phonology was to present criteria for relevance, criteria of selection, to apply to the mass of unorganized phonetic data. Now we seem to have had some fifteen years of something different: phonology as speculative phonetics, and we

are now not so much imposing criteria of relevance on phonetic research as asking the phonetician to provide us with criteria to decide amongst different formal systems of phonology. Amongst these criteria will be the old functional phonology which is now, as it were, part of the surface data.

Kenneth L. Pike: It is not possible to separate phonology from grammar, from lexicon, from meaning. We must have a trihierarchical structure: phonology, grammar, and meaning. But in each of the hierarchies there are thresholds. - No mathematical system of any complexity can be treated as consistent by looking at the data inside itself. Something external must be used. That which I use from outside the formal system, to make it relevant, is meaning and behavioral impact.

Hans Basbøll: I want to stress once more that if my report is to be read as a status report on phonology, it should be read in connection with the contributions to the symposia.

Stephen Anderson: Perhaps we can all agree that the fundamental problem for phonologists is the exploration of what can constitute the sound pattern of a language. Ultimately we all have to make our own choice about what is the most productive way to go about this investigation, and I think it is unlikely that there are determinate answers to the sorts of opposition questions that have been posed.

THE RELATIONS BETWEEN AREA FUNCTIONS AND THE ACOUSTICAL SIGNAL

Gunnar Fant, Department of Speech Communication, Royal Institute of Technology, S-10044 Stockholm, Sweden

Chairpersons: Wiktor Jassem and Kenneth N. Stevens

Introduction

The topic of this paper is to discuss how configurations, shapes, and detailed outlines of the vocal tract cavity system influence the acoustic signal and the reverse, how to predict vocal tract resonator dimensions from speech wave data. As far as the direct transform is concerned, this is a re-visit to my old field of acoustic theory of speech production.

What progress have we had in vocal tract modeling and associated acoustic theory of speech production during the last 20 years? My impression is that the large activity emanating from groups engaged in speech production theory and in signal processing has not been paralleled by a corresponding effort at the articulatory phonetics end. Very little original data on area functions have accumulated. The Fant (1960) Russian vowels have almost been overexploited. Our consonant models are still rather primitive and we lack reliable data on details of the vocal tract as well as of essential differences between males and females and of the development of the vocal tract with age.

The slow pace in articulatory studies is of course related to the hesitance in exposing subjects to X-ray radiation. Much hope was directed to the transformational mathematics for deriving area functions from speech wave data. These techniques have as yet failed to provide us with a new reference material. The so-called inverse transform generates "pseudo-area functions" that can be translated back to high quality synthetic speech but which remain fictional in the sense that they do not necessarily resemble natural area functions. Their validity is restricted to non-nasal, non-constricted articulations and even so, they at the best retain some major aspects of the area function shape rather than its exact dimensions. However, some improvements could be made if more representative acoustic models than LPC analysis are considered.

Once a vocal tract model has been set up it can be used, not only for studying articulation-to-speech wave transformations, but also for a reverse mapping of articulations and area functions to fit specific speech wave data. These analysis-by-synthesis re-

mapping techniques, as well as perturbation theory for the study of the consequences of incremental changes in area functions or of the inverse process, are useful for gaining insight in the functional aspect of a model. However, without access to fresh articulatory data the investigator easily gets preoccupied with his basic model and the constraints he has chosen.

The slow advance we have had in developing high quality synthesis from articulatory models is in part related to our lack of reliable physiological data, especially with respect to consonants, in part to the difficulty involved in modeling all relevant factors in the acoustic production process. The most successful attempt to construct a complete system is that of Flanagan et al. (1975) at Bell Laboratories. A variety of studies at KTH in Stockholm and at other places have contributed to our insight in special aspects of the production process such as the influence of cavity wall impedance, glottal and subglottal impedance, nasal cavity system, source filter interaction, and formant damping.

From area function to the acoustic signal

The acoustic signal or, in other words, the speech wave is the product of a source and a filtering process. The most common approach is to disregard the source and relate a vocal tract area function to a corresponding formant pattern only, i.e. a set of formant frequencies F_1 F_2 F_3 F_4, etc. This correspondence is illustrated by Fig. 1. I shall not go into the mathematics of the wave equations and the equivalent circuit theory. Instead I will attempt to develop a perspective around some basic models and current problems.

To derive an area function from X-ray data on vocal tract dimensions is by no means a straightforward procedure, see Fant (1960; 1965) and Lindblom and Sundberg (1969).

The estimation of cross-sectional shapes and dimensions in planes perpendicular to the central pathway of propagation through the vocal tract has to rely on crude conventions and involves uncertainties, e.g. with respect to variations with articulation and for different types of subjects. The lack of basic data is especially apparent for female and child speech and for consonants, e.g. laterals and nasals. In spite of the accessibility of the speech wave to quantitative analysis there is a similar lack of reference data concerning the acoustic correlates. Most studies have been concerned with male speech and vowels.

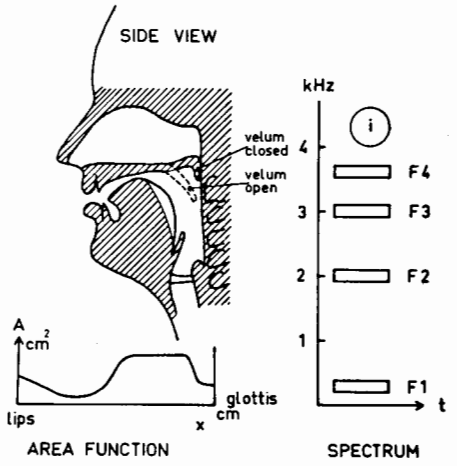


Figure 1. Principle illustration of vocal tract sagittal view with area function and corresponding resonance frequency pattern.

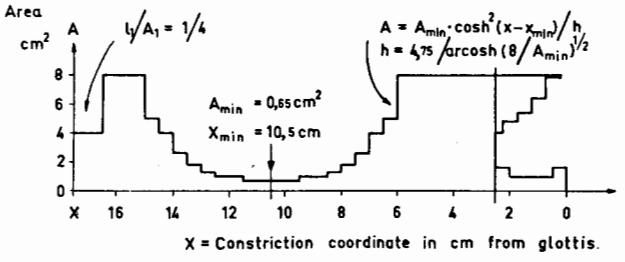


Figure 2. Three-parameter vocal tract model (Fant, 1960).

A specification of an area function as a more or less continuous graph of cross-sectional area from the glottis to the lips allows detailed calculations of the acoustic response but is not practical for systematic descriptions. A data reduction in terms of parametric models brings out the acoustically relevant aspects. The three-parameter models of Stevens and House (1955) and Fant (1960) differ somewhat in the details but have the same set of parameters, the place of minimum cross-sectional area of the tongue section, the area at this coordinate, and the length over area ratio l_0/A_0 of the lip section.

My model is shown in Fig. 2. The shunting sinus piriformis cavity around the outlet of the larynx tube was a constant feature in my model. A weakness is that it is not reduced in volume for back vowels which does not allow F_1 to reach a sufficiently high value for [ɑ]. Fig. 3 shows the variation of the F-pattern with the place of tongue constriction. This is a well established graph which retains basic patterns such as the rise of F_2 with advance of the tongue constriction from back to front up to an optimal place at a midpalatal location after which F_2 drops again. A limitation of the parameter range to a region bounded by [ɑ], [u], and [i] as proposed in several articulatory models, e.g. Lindblom and Sundberg (1969), would exclude the standard Swedish pronunciation of the vowel [ʉ] which, contrary to traditional classifications, has a constriction somewhat anterior to that of [i] (Fant, 1973).

The constriction coordinate is an acoustically more relevant classifier than the "highest point of the tongue" of classical phonetics. Most stressed vowels have a definite "place of articulation" as evidenced by a region of minimum cross-sectional area which we may exemplify by [i], [u], [o], [ɑ] ending with a variant of [æ] with major narrowing just above the glottis (Fant, 1960). On the other hand, it may be argued that the traditional classification in terms of tongue locations and related parameters belongs to a production stage one step higher up than area functions and could be directly related to formant patterns.

The [ɑ] and [i] vowels are polar opposites, the [i] vowel requiring a wide pharynx and narrowed mouth, whilst the opposite is true of [ɑ] type vowels. A production of a vowel [u] requires a double resonator configuration with a narrow lip opening to ensure

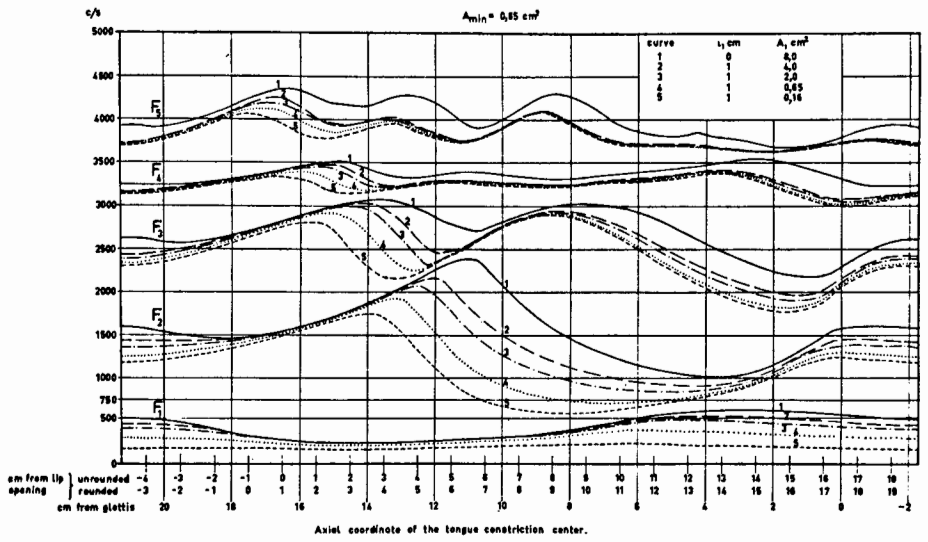


Figure 3. F-pattern variation with constriction coordinate x_c at different sets of lip parameter l_1/A_1 at constant constriction area A_{min}. The constriction coordinate is zero at the glottis.

a low F_1 and a narrow constriction between the two major cavities as a correlate of a low F_2. These shape aspects are brought out in the stylized area functions of Fig. 4. A basic issue in acoustic phonetics is that it is not possible to produce these vowels without retaining the major shape aspects of the area functions. To this extent area functions are predictable from the acoustic signal as will be discussed in greater detail in a later section. Peter Ladefoged would back me up here with his competence of transforming phonetic qualities to equivalent resonator configurations.

Another basic issue is that the vocal tract filtering is determined by the location of formants only and that the spectrum envelope between peaks cannot contain any other irregularities than those originating from the source function. Minor irregularities in the outline of the area function may have some influence on formant locations but will not give rise to irregularities in the spectrum envelope. This is not evident without an insight in the mathematical constraints imposed by acoustic theory. It is related to the one-dimensional wave propagation, wavelengths generally being short compared to vocal tract cross dimensions. Systematic perturbations of vocal tract area functions will be discussed in a later section.

Highly simplified area functions of fricatives (or corresponding stops) and their filtering functions are shown in Fig. 5. As discussed by Fant (1960), the "compact" sibilant [ʃ] or the stop [k] has a definite cavity in front of the major constrictions which accounts for a central dominance of the spectrum, usually a single formant, if the cavity is abruptly terminated by the constriction. The [s] or [t] has a narrow channel of a few centimeters length behind the source which may combine with a small front cavity to produce resonances above 4000 Hz which build up a high-pass filtering. The [f] or [p] has no significant resonance in its closed state.

In general, the cavities behind the source do not influence the spectrum much, provided that the consonantal constriction is effective. Resonances of the back cavities may appear if the constriction tapers off gradually as in palatals or if a palatal tongue articulation builds up a supporting constriction behind the lips. Back cavity resonances combine with and are cancelled by spectral zeroes at complete closure but move away from their

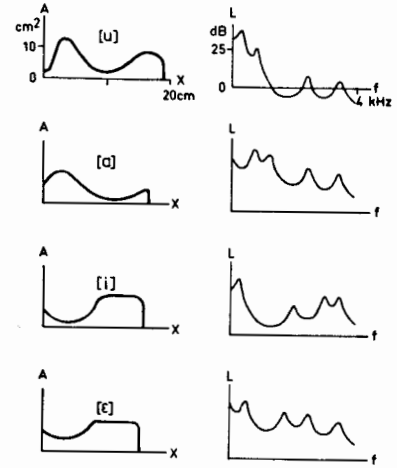


Figure 4. Stylized area functions and corresponding spectrum envelopes of [u] [ɑ] [i] and [ɛ]. The constriction coordinate is zero at the lips.

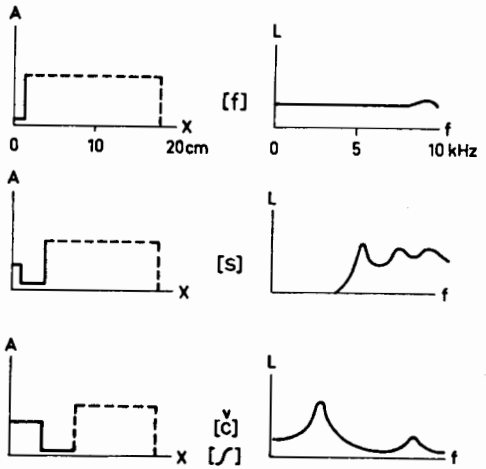


Figure 5. Stylized area functions and corresponding spectra of three basic consonant categories. The constriction coordinate is zero at the lips.

zero mates during release and are then more or less free to appear. In Fig. 6 we can study measured and calculated spectra of [k] and a palatalized [p'] (Fant, 1960). The labial burst spectrum contains peaks at around 2-3 kHz but has a free spectral minimum at 1400 Hz. In contrast, the [k] spectrum has a single formant peak around 1400 Hz. It is interesting to note that the calculations from the area function data back up the measured spectra. We need more studies of this type.

Vocal tract boundary constraints and dynamics

The simplified static models relating a single area function without parallel branches to a set of formant frequencies have obvious limitations. On a higher level of ambition we must include proper boundary conditions such as radiation load and a finite coupling to the subglottal and nasal systems. In order to predict formant bandwidths we must consider the energy loss during an oscillatory cycle of a formant associated with "loss elements" on the surface of the vocal tract resonator system and other dissipative elements (Fant and Pauli, 1975). Source functions must be defined with respect to place of insertion in the vocal tract, their spectrum or waveform, and the degree of coupling to other parts of the system (Stevens, 1971). In addition, these properties are highly time variable within a voice fundamental period (Fant, 1979) and within intervals of transition from various states of the glottis or of other terminations of the vocal tract. Rapid opening and closing gestures pose specific problems in relating area functions to acoustic data. In a proper analysis of connected speech we need two sets of acoustic variables: the continuous variations of the F-pattern as a correlate of the continuous movements of the articulators and the often abruptly varying patterns of spectral energy distributions associated with discrete events of production.

The acoustic production model of Fig. 7 may serve as a starting point for a brief discussion of these problems. First of all, we should note an important element in converting area functions to a filter function. The walls of the vocal tract are not rigid. They may expand during a voiced occlusion as represented by the element C_w in the equivalent circuit of a small slice of the area function, Fig. 8, and they have a finite mass L_w which adds to the tuning of vocal resonances and which dominates the impedance of

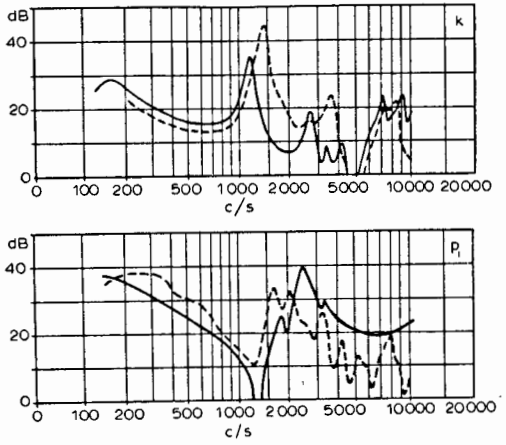


Figure 6. Calculated (solid line) and measured (broken line) stop release spectra of a velar [k] and a palatalized [p']. The minimum in [p'] at 1400 Hz is a free zero in the sub-lip impedance whilst the main formant of [k] is a mouth cavity formant. After Fant (1960).

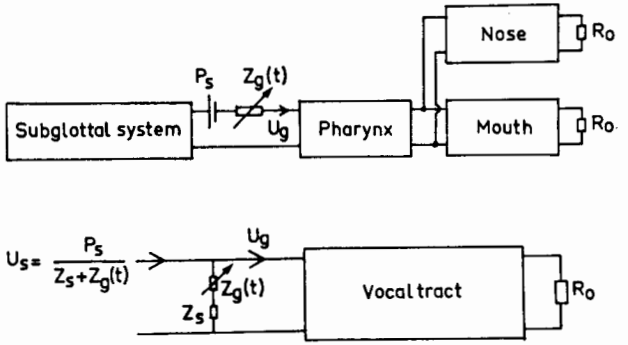


Figure 7. Block diagram of the production of voiced sounds illustrating the principle difference between an ideal source U_s and a glottal flow U_g.

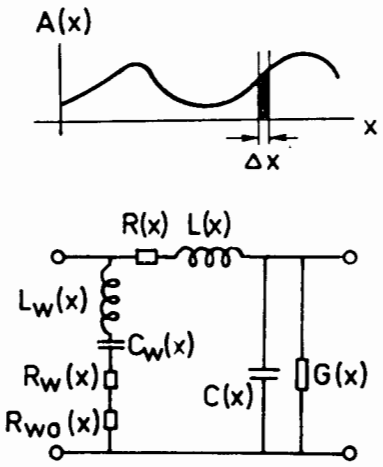


Figure 8. Lumped constant approximation of a small slice of the area function.

the shunting branch at frequencies above 40 Hz. A small fraction of sound is radiated externally from the outside of the head through R_{wo}. It is negligible except as a constituent of the voice bar of a voiced occlusion.

Disregarding the cavity wall mass element L_w, calculations would provide $F_1 = 0$ for an area function starting and ending with complete closure. The finite F_1 of around 150-250 Hz found in the spectrogram of the voiced occlusion is determined by the resonance of the entire air volume compliance in the tract with the total lumped cavity wall mass shunt. This resonance can easily be measured acoustically (Fant et al., 1976) and amounts to $F_{1w} = 190$ Hz with a bandwidth of $B_{1w} = 75$ Hz, typically for a male voice, and around 20% higher for females. The wall mass element L_w is thus an important constituent in calculating F_1 from the area function. The procedure is to start out with a derivation of an ideal F_{1i} without mass shunt and add a correction factor

$$F_1 = F_{1i}(1 + F_{1w}^2/F_{1i}^2)^{1/2} \qquad (1)$$

The distribution of the wall impedance along the vocal tract and its dependence on particular articulations are not known. The experiments of Fant et al. (1976) suggest that regions around the larynx and the lips are especially important. Experiments by Ishizaka et al. (1975) provide data of the same order of magnitude but have not revealed conclusive distribution patterns.

The resistive component R_w in the cavity wall branch determines a major part of the bandwidth B_1 of low F_1 formants. The resistive part of the radiation load which is proportional to frequency squared is the essential bandwidth determinant of resonances above 1000 Hz originating from an open front resonator. Internal surface losses from friction and heat conduction enter through the elements R and G in Fig. 8. They are proportional to the half power of frequency and to the inverse of the cross-sectional area. A detailed analysis of formant bandwidths and their origin appears in Fant (1972), Fant and Pauli (1975), and Wakita and Fant (1978).

The time variable glottal impedance accounts for variations of formant frequencies and bandwidths within a voice fundamental period (Flanagan, 1965). A more detailed analysis of glottal damping requires a reconsideration of the process of voice generation (Fant, 1979) and adoption of perceptual criteria for deriving equivalent mean values (Fant and Liljencrants, 1979). The main excitation of the vocal tract occurs at the instant of interruption of glottal flow by glottal closure. At this instance, damped oscillations are evoked and subjected to the damping from supraglottal loss elements.

When the glottis opens for the next flow pulse the vocal tract becomes loaded by the time variable glottal plus subglottal impedance. Providing a resonance mode is much dependent on the part of the area function immediately above the glottis, the glottal damping becomes severe. This is especially apparent if the lower pharynx is narrowed thus facilitating an impedance match between the cavity system and the glottal resistance. A complete extinction of the formant oscillation in the glottal open interval may result. This is typical of F1 of the vowel [a] produced at low or moderate voice effort by a male subject.

In general most of the energy excited during a voice fundamental period is lost during the timespan of the following period. Since glottal resistance decreases with lowered transglottal pressure the damping effect is especially apparent at weak voice levels. The mean glottal bandwidth in normal voice production is of the order of 0-100 Hz with 20 Hz as a typical value for male medium intensity phonation.

It is apparent that any model of voice production which adopts the actual flow through the glottis as the primary source will create problems. With this convention, which happens to apply to inverse filtering techniques, the source attains components of formant oscillations and becomes dependent of the vocal tract area function (Mrayati and Guérin, 1976). Their approach is intended to define a proper source for a formant synthesizer.

A different approach more suited for production models is to incorporate the combined glottal and subglottal impedance as a termination paralleling the input end of the tract and to define the source as the flow through the glottis which would have occurred with the input to the vocal tract short circuited. This representation adopted by Fant (1960) preserves a realistic definition of the vocal tract transfer function but fails to take into account source modifications due to aerodynamic losses in supraglottal constrictions. In the transition from a vowel to a voiced consonant there is generally some loss of transglottal pressure which reduces the excitation strength of the voice source.

The interplay of glottal and supraglottal sources associated with articulatory narrowing and release becomes an important part of a dynamically oriented theory of predicting acoustic signals from area functions (Stevens, 1971).

What about the subglottal system? How does it influence speech? In normal voice production the influence appears to be small. As long as the glottal opening is small and the flow velocity high, the glottis impedance becomes high compared to the subglottal impedance. Unless there is a constant leakage bypassing the vibrating part of the glottis, the subglottal system should have a minor influence only.

This reasoning is concerned with the modification of the supraglottal formants only. At the instance of flow interruption when the glottis closes there is a simultaneous excitation of resonances in the trachea and other parts of the subglottal system. Potential frequencies are 600, 1250, and 2150 Hz for a male voice (Fant et al., 1972). The transmission losses associated with the penetration of these components through the walls of the trachea and the chest to externally radiated sound appear to be sufficiently high to rule out any significance, but this remains to be proved.

As shown by Fant et al. (1972), subglottal formants may occasionally be seen in spectra from aspirated sound segments, e.g. in the release phase of unvoiced stops. "F1-cutback" in the first part of the voiced interval after release, which appears as a relative delay in onset of F1 compared to F2 and higher formants, may be explained as an instance of excessive F1 damping through an incompletely closing glottis. The upper formants are less dependent on the glottal termination and thus less affected. This relative weakening of F1 is a filtering effect, whilst the relative weakness of F1 in a preceding unvoiced, aspirated segment is also a matter of low source energy in the F1 region. The F1 intensity reduction is also seen in the terminating periods of a vowel before the occlusion of an unvoiced stop (pre-occlusion aspiration).

Nasalization and aspiration have similar effects on F1. In nasalized sounds the F1 intensity is typically reduced by a spectral zero (Fant, 1960; Fujimura and Lindqvist, 1971). The nasal model of Fant (1960) produces too high values of the lowest nasal pole. The possible occurrence of several low frequency pole-zero pairs is made plausible by the study of Lindqvist and Sundberg (1972). More anatomical and acoustic data are needed.

In connection with the voice source studies of Fant (1979) it has been noted that the spectral maximum often seen below F_1 in vowels is a voice source characteristic, which becomes especially enhanced in contrast to a weak F1 in nasalized or aspirated, voiced segments. This is especially apparent in a time domain study. Another way of expressing this finding is to say that nasal sounds retain more source characteristics than non-nasal sounds.

If an area function is subjected to a substantial change in a very short time, one may expect some deviations from the linear stationary behavior. Point-by-point calculations of resonance frequencies are still valid but additional bandwidth terms enter which may be positive or negative. A rapid opening of a constriction is accordingly associated with a negative bandwidth component and a rapid closure with a positive bandwidth component. The analysis is simple. Consider a flow U(t) through an acoustic inductance $L(t) = \rho l/A(t)$. The pressure drop is:

$$P(t) = \frac{d}{dt}\,[L(t)U(t)] = L'U + LU' \qquad (2)$$

L' = dL/dt apparently has the dimension of a resistance R_d

$$R_d = \frac{dL}{dt} = \frac{-A'(t)\rho l}{A^2(t)} \qquad (3)$$

In a single resonator system the bandwidth component associated with a resistance R in series with an inductance L is simply $R/2\pi L$.

Accordingly, the bandwidth associated with R_d is

$$B_d = \frac{-A'(t)}{2\pi A(t)} \qquad (4)$$

which implies a bandwidth component of opposite sign to that of the rate of change of the area. Fig. 9 illustrates the temporal course of the bandwidth when a resonator of volume 100 cm^3 is coupled to a neck of length 4 cm and a cross-sectional area A(t) varying exponentially from closure to complete opening of 2 cm^2 with a time constant of 10 milliseconds.

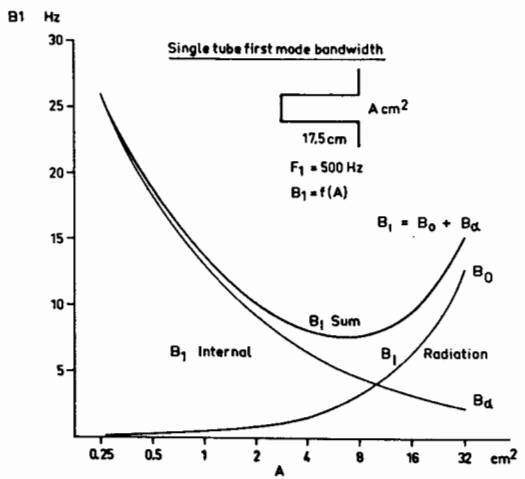


Figure 9. Resonator outlet area A, resonance frequency F, and total bandwidth B as a function of time during an exponential release with a time constant of 10 milliseconds. B_d is the negative dynamic component of the bandwidth.

The time varying negative bandwidth overrides the frictional bandwidth components up to 8 milliseconds after release which could tend to increase the amplitude of the oscillation during that period. However - in the speech case there enter additional positive bandwidth components related to flow dependent resistance and to cavity wall losses and possibly also glottal losses which tend to reduce the importance of the negative terms. In a detailed analysis of the glottis resistance the dynamics calls for some decrease of glottal resistance in the rising branch of the glottal pulse and an increase in the falling branch, as noted by Guérin et al. (1975). Except for the analysis above, a proper evaluation of the practical significance has to my knowledge not been performed. The most detailed thesis on the theoretical aspects is that of Jospa (1975). I feel that dynamic effects are of academic rather than practical significance. Of greater importance is probably the mere fact that a rapid transition of a formant creates a special perceptual "chirp" effect.

Perturbation theory and vocal tract scaling

Perturbation theory describes how each resonance frequency, F_1 F_2 F_3, etc., varies with an incremental change of the area function A(x) at a coordinate x and allows for a linear summation of shifts from perturbations over the entire area function. The relative frequency shift $\delta F/F$ caused by a perturbation $\delta A(x)/A(x)$ is referred to as a "sensitivity function". We may also define a perturbation $\delta\Delta x/\Delta x$ of the minimal length unit Δx of the area function which will produce local expansions and contractions of the resonator system. It has been shown by Fant (1975b), Fant and Pauli (1975) that the sensitivity function for area perturbations of any A(x) is equal to the distribution with respect to x of the difference $E_{kx}-E_{px}$ between the kinetic energy $E_{kx} = \frac{1}{2}L(x)U^2(x)$ and the potential energy $E_{px} = \frac{1}{2}C(x)P^2(x)$ normalized by the totally stored energy in the system.

Fig. 10 from Schroeder (1967) illustrates perturbations of a single tube resonator by changes in the area function derived from sinusoidal functions. These have been chosen to influence F_1 only (a), none of the formants (b), and F_2 only (c). The middle case is of special interest. There exists an infinite number of small perturbations applied symmetrically with respect to the midpoint of the single tube, which will have almost no influence

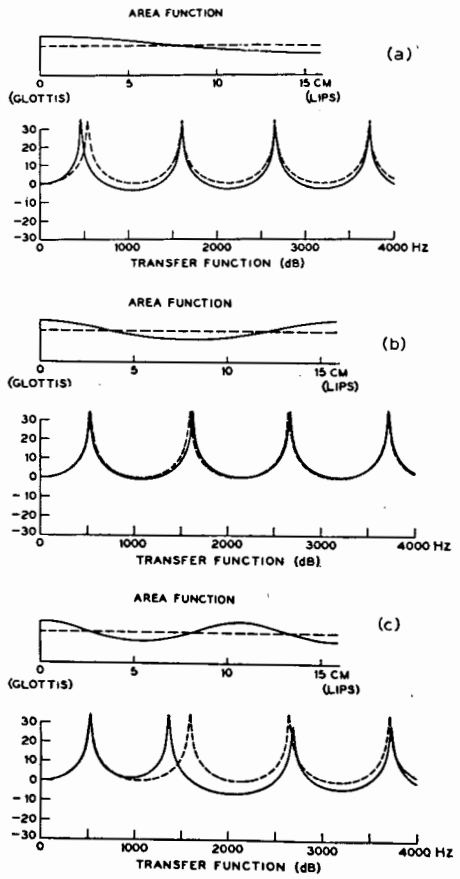


Figure 10. Perturbations of the single tube area function affecting F_1 only (a), almost no influence (b), and F_2 only (c) (after Schroeder, 1967).

on the formant pattern. In the general case of an arbitrary area function the rule of symmetry is upset (Heinz, 1967) but there still exists a tendency of compensatory interaction between front and back parts (Öhman and Zetterlund, 1975).

Sensitivity function for area perturbations of my six Russian vowels are shown in Fig. 11. This chart is useful as a reference for general use. Given the relative amount of area change, the corresponding relative frequency shift $\delta F_n/F_n$ is proportional to the product of $\frac{\delta A(x)}{A(x)}$ and the amplitude of the sensitivity function, $E_{kx}-E_{px}$. As an example we may note that F_1 of the vowel [u] rises with increasing area at the lips, i.e. decreases with increasing degree of narrowing and that narrowing the tongue constriction of [u] causes F_2 to fall and F_3 to rise. A narrowing of the outlet of the larynx tube will apparently have the effect of tuning F_4 to a lower frequency.

With the area function sampled at intervals of Δx, e.g. $\Delta x = 0.5$ centimeter for practical use, we may ask what happens if we increase Δx at the coordinate x by the amount $\delta\Delta x$. The local expansion thus introduced causes a frequency shift $\delta F_n/F_n$, which is proportional to $-\delta(x)/1+\delta(x)$ and to $(E_{kx}+E_{px})$ of resonance n.

The distribution of $(E_{kx}+E_{px})$ is uniform for a single tube resonator. The effect of a length increase is obviously the same irrespective of where along the x-axis the tube is lengthened. An overall increase of the length by, say $\delta(x) = 0.2$, causes a shift of all resonance by a factor $-0.2/(1+0.2) = -0.17$. The same calculation performed directly from the resonance formula $(2n-1)c/4l_t$, where l_t is the total length and c=35300 cm/s is the velocity of sound, would provide the same answer, i.e. a frequency ratio of $1/(1+0.2)=0.83$.

The distribution $E_{kx}+E_{px}$ along the vocal tract is also a measure of the relative dependence of the particular resonance mode on various parts of the area function. This is the best definition we have of "formant-cavity" affiliations. From Fig. 12 we may thus conclude that most of the energy of the second formant of [ɪ] is stored in the pharynx, whilst the third formant of [i] "belongs to" the front part of the system. F3 of the back vowels [u] [o] and [ɑ] are associated with a central part of the tract, and F4 of all vowels has a substantial peak of energy located in

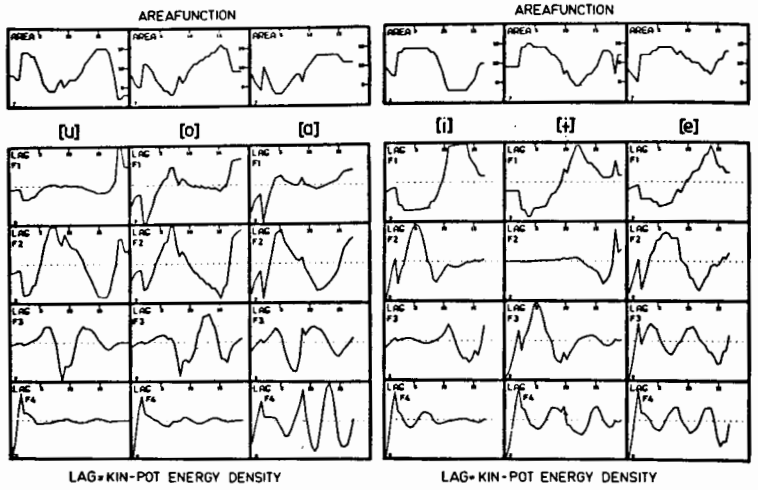


Figure 11. Sensitivity functions for area perturbations of the six Russian vowels (Fant, 1960). From Fant (1975b). The constriction coordinate is zero at the glottis.

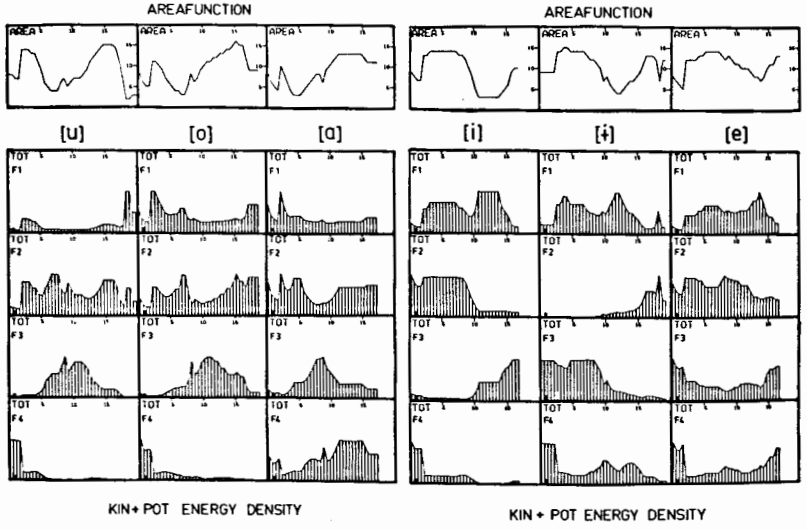


Figure 12. Sensitivity functions for length perturbations of the six Russian vowels (Fant, 1960). From Fant (1975b). The constriction coordinate is zero at the glottis.

the larynx tube. Expanding the length of the pharynx will have a large effect on F_2 of [i] and a small effect only on F_3 and vice versa for a length expansion of the mouth cavity. This analysis would apply to the relatively short pharynx of females compared to males.

If a perturbation of the entire area function is expressed as a function of as many parameters as there are formants, it is possible to calculate the change in area function from one F-pattern to another (Fant and Pauli, 1975). This indirect technique has been used by Mrayati et al. (1976) for deriving plausible area functions for French vowels on the basis of their deviation from my reference Russian vowels. This procedure must be administered in steps of incremental size with a recalculation of the sensitivity function after each major step. It may involve length as well as area perturbations.

In practice, when aiming at direct transforms only, it may be easier to resort to a direct calculation of the response of the perturbed area functions than to derive it from the energy distributions. The perturbation formulas and especially their energy based derivations are more useful for principal problems of vocal tract scaling or for gaining an approximate answer to a problem without consulting a computer program.

The area functions of male and female articulations of the Swedish vowels [i] and [u] and corresponding computed resonance mode pattern in Fig. 13 may serve to illustrate some findings and problems. The data are derived from tomographic studies in Stockholm many years ago in connection with the study of Fant (1965; 1966) and were published in Fant (1975a; 1976). It is seen that in spite of the larger average spacing of formants in the female F-pattern related to the shorter overall vocal tract length, the female F_1 and F_2 of [u] and the F_3 of [i] are close to those of the male. This is an average trend earlier reported by Fant (1975a), see Fig. 14. Differences in perceptually important formants may thus be minimized by compensations in terms of place of articulation and in the extent of the area function narrowing. Such compensations are not possible for all formants and cannot be achieved in more open articulations. The great difference in F_2 of [i] is in part conditioned by the relatively short female pharynx but can in part be ascribed to the retracted place of

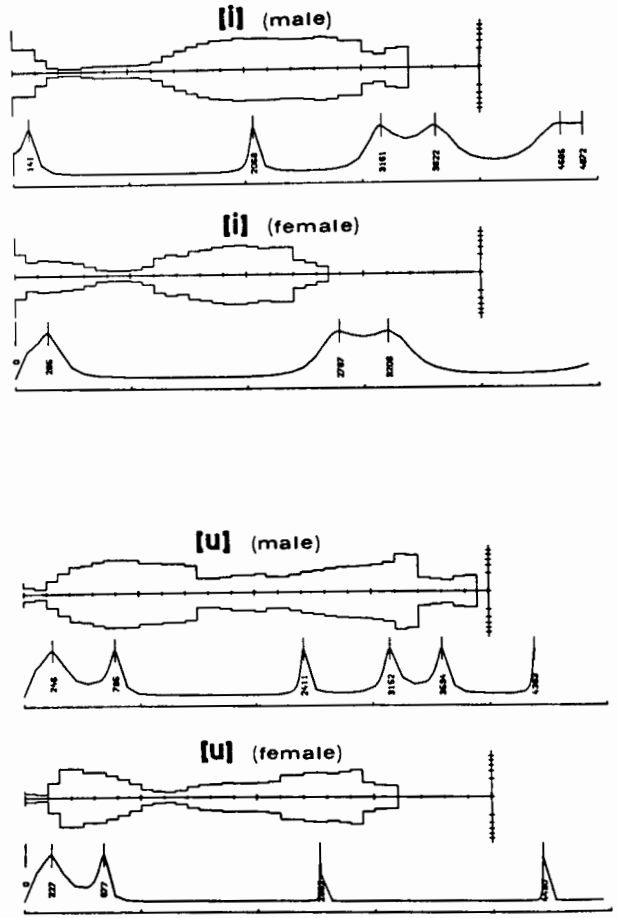


Figure 13. Male and female vocal tracts (equivalent tube representation) and corresponding F-patterns from the tomographic studies of Fant (1965).

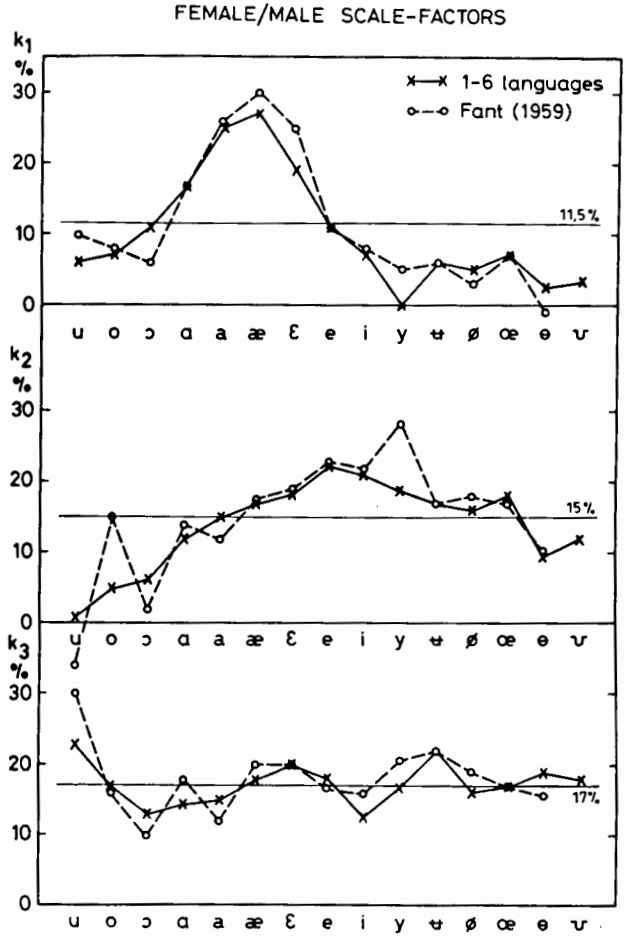


Figure 14. Female/male scale factor variation with vowel and the particular formant (Fant, 1975a).

articulation. It is also disputable whether this particular female articulation serves to ensure an acceptable [i] or whether there is a dialectal trend towards [ɪ]. Also, it is to be noted that X-ray tomography may impede the naturalness of articulations because of the abnormal head position required.

Much remains to be studied concerning how the vocal tract area functions of males, females, and children are scaled in actual speech and what kind of compensation occurs for minimizing perceptual differences or maybe the reverse, to mark contrasts between age and sex groups.

The lack of reference data on area functions is severe and the attempts to overcome this lack by means of area function scaling performed by Nordström (1975) were not conclusive except to support the general issue that the vowel and formant specific female-male differences, documented by Fant (1975a), Fig. 14, do not always come out as a result of the particular scaling assumed. The agreement was good for F_3 and fair for F_2 and rather bad for F_1. The predictability of F_3 is expected in view of the high dependency of F_3 on length dimensions.

A weakness in the Nordström study is that his [æ] and [ɛ] vowel area functions were interpolated from the Russian [ɑ] and [e] vowel and accordingly attain a centralized quality not representative of the [ə] and [æ] category vowels which normally display a very large female-to-male F_1 ratio, see Fig. 14.

It is interesting to note that the non-uniform differences between females and males are paralleled by similar patterns comparing tenor and bass male singers. These vowel and formant specific trends are not only the automatic consequence of different anatomical scalings but also reveal compensations according to criteria that are not very well understood yet. A promising project on vocal tract modeling from anatomical data, now carried out at MIT (Goldstein, 1979), should provide us with fresh insight in female, male, and child differences.

From Goldstein's still unpublished graphs of vocal tract outlines I have noted that the length of the pharynx measured from the glottis to the roof of the soft palate grows from 3.3 cm in the newborn child to 7.6 cm for the female aged 21 and 10 cm for the male aged 21. The length of the mouth measured from the back wall of the upper pharynx to the front teeth (alveolar ridge for the newborn infant) grows from 5.5 cm for the newborn infant to 8 cm for the female of 21 and 8.5 cm for the male of 21. The tendency of relatively small variations of mouth cavity length with sex and age is more apparent than anticipated from earlier studies and would tend to minimize the range of "mouth cavity formant frequencies". The radical variations in relative pharynx length suggest that the relative role of front and back parts of the vocal tract could be reversed for a small child, i.e. that F_2 of the vowel [i] would be a front cavity formant, whilst F_3 is more dependent on the shorter back cavity. When front and back cavities are of more equal length, the dependency is divided and the F_3/F_2 ratio smaller than for males, which is typical of females or children of an intermediate age.

The inverse transform

As noted already in the introduction, there has been a substantial amount of theoretical work directed towards the derivation of area functions from speech wave data. In practice, however, these techniques are limited to non-nasal, non-obstructed vocal productions and the accuracy has not been great enough to warrant their use in speech research as a substitute for cineradiographic techniques. In the following section I shall attempt to comment on some of the main issues and problems. The usual technique, e.g. Wakita (1973), is to start out with a linear prediction (LPC) analysis of the speech wave to derive the reflection coefficients which describe the analog complex resonator. The success of this method is dependent on how well the losses in the vocal tract are taken into account. Till now the assumptions concerning losses have been either incomplete or unrealistic. Also the processing requires that the source function be eliminated in a preprocessing by a suitable deemphasis or by limiting the analysis to the glottal closed period. In spite of these difficulties the area functions derived by Wakita (1973; 1979) preserve gross features.

In general, a set of formant frequencies can be produced from an infinite number of different resonators of different length. We know of many compensatory transformations, such as a symmetrical perturbation of the single-tube resonator. However, if we measure the input impedance at the lips (Schroeder, 1967) or calculate formant bandwidths, we may avoid the ambiguities. A tech-

nique for handling tubes with side branches has been proposed by Ishizaki (1975).

According to Wakita (1979), the linear prediction method is capable of deriving an area function quantized into successive sections of equal and predetermined length providing the LPC analysis secures an analysis equivalent to M formants specified in terms of frequency and bandwidth.

An estimation of the total length and of the area scale factor require additional analysis data. An incorrect length estimate automatically generates compensatory changes in the area function which may be appreciable.

LPC analysis is a simple and powerful method of analysis but it fails in naturalness of representing the production process and as such is a poor substitute for a lossy transmission line representation. With the fresh eyes of a non-expert on the inverse transform, I would attempt to make the following suggestions. One is that M formants with associated bandwidths could have a greater predictive power than noted by Wakita. The area scale factor could be included in addition to the 2M relative areas of his model. In general, with reservation for possible uniqueness problems, 2M formant parameters - including bandwidths but not necessarily as many bandwidths as frequencies - would suffice for predicting 2M independent area function parameters.

Thus, adding one more formant frequency to the M pairs of frequencies and bandwidths would suffice for estimating the total length of the 2M system. Alternatively, from the 2M formant measures, we could derive a model quantized into M equivalent tubes each specified by cross-sectional area and specific length, thus also predicting the total length, Fig. 15. The rationale for this reasoning is that all losses in the transmission line analogs are unique functions of the area and length dimensions. One could also design a three-parameter model of the vocal tract as in Fig. 15 with a constant larynx tube. The four parameters (lip parameter A_1/l_1, x_c and A_c, and the total length) would hopefully be predictable from a specification of F_1, F_2, and F_3 and a bandwidth, say B_3, which appears to be more discriminating than B_1 and B_2. If we omit the total length and sacrifice the bandwidth, we have approached the articulatory modeling of Ladefoged et al. (1978)

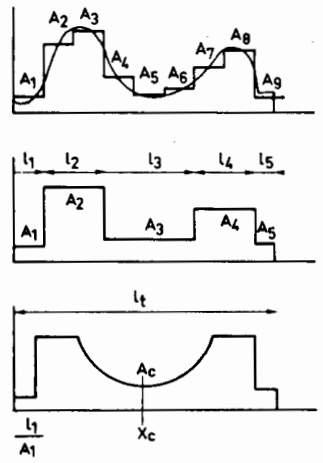


Figure 15. Continuous area function approximated by a constant larynx tube and 8 sections of equal length (top), by 4 sections of variable length and area (middle), and by a three-parameter model extended to include the total length (bottom). The constriction coordinate is zero at the lips.

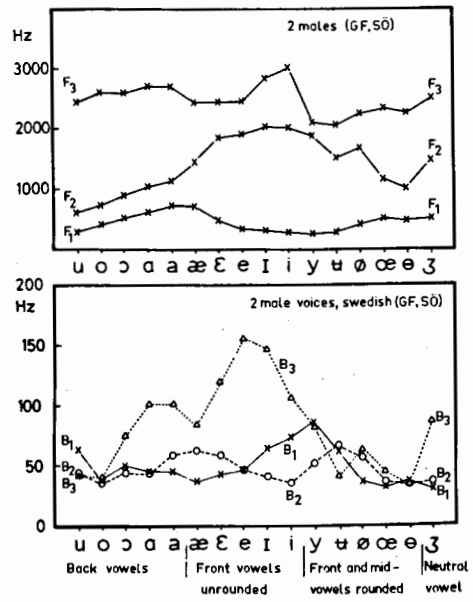


Figure 16. Frequency and bandwidth patterns of Swedish vowels (Fant, 1972).

which is based on correlational methods for deriving three articulatory parameters from F_1, F_2, and F_3.

In general, bandwidths have less predictive power than frequencies. They are to some extent predictable from formant frequencies (Fant, 1972), Fig. 16. Furthermore, bandwidths vary with speaker, voice effort, and laryngeal articulations and are inherently difficult to measure.

Still, I do not want to rule out the use of bandwidths. The following examples may serve to illustrate their predictive power and limitations. First, a test of the uniqueness in predicting 2M area function parameters from 2M formant data. Take the simple case of M=1 which implies a single tube resonator. What are the length and cross-sectional area of a tube with a specified first resonance frequency and bandwidth? The length is immediately given by $F_1=c/4l$. As shown in Fig. 17 the area is a single-valued function of bandwidth providing only one loss element is postulated (as in LPC analysis). If we include both the internal surface losses of a hard-walled tube and the radiation resistance, the bandwidth versus area attains a minimum at 10 cm^2 and there are two alternative areas that fit the same bandwidth. The higher value could possibly be ruled out as being outside the possible range of human articulation. Similar ambiguities could also be expected in a more complex lossy transmission line model, as pointed out by Atal et al. (1978). However, one should note that their treatment of the invariance problem is not quite fair. They introduce more articulatory parameters than acoustic descriptors which obviously exaggerate the ambiguities. Next consider a two-tube approximation of the vocal tract, Fig. 18 (A), with a back tube of length 8 cm and area 8 cm^2 and a front tube of length 6 cm and cross-sectional area 1 cm^2. The formant frequency pattern of F_1=275 Hz, F_2=2132 Hz, F_3=2998 Hz, F_4=4412 Hz and all higher formants is exactly the same as that of a two-tube system with the same areas but the lengths reversed, i.e. a front tube of length 8 cm and a back tube of length 6 cm (Fig. 18 B). This length ambiguity rule is apparent from the expression for resonance conditions

$$\frac{A_2}{A_1} \operatorname{tg} \frac{\omega l_1}{c} \times \operatorname{tg} \frac{\omega l_2}{c} = 1 \qquad (5)$$

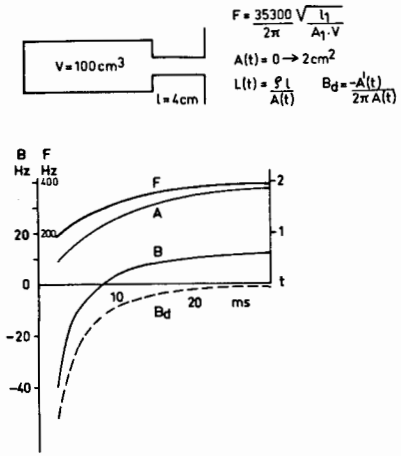


Figure 17. Bandwidth versus area of a single tube resonator taking into account internal losses and radiation load losses.

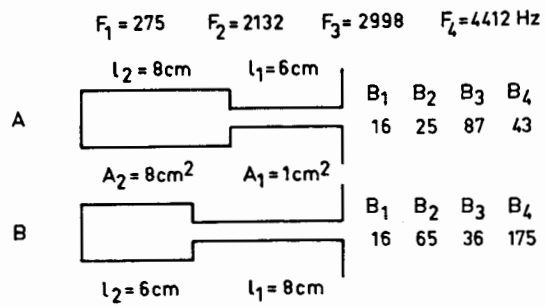


Figure 18. Two twin-tube resonators that provide the same F-pattern appropriate for the vowel [ɪ], differing in terms of bandwidths.

If bandwidths are calculated taking into account both the interior surface losses and the radiation resistance by formulas given by Fant (1960), we find that B_2 and B_4 of Fig. 18 (A) are relatively low compared to B_3. In Fig. 18 (B), B_2 and B_4 are large compared to B_3. The different bandwidth patterns resolve the ambiguity. The physical explanation is that F2 and F4 of the first model are essentially determined by the back cavity and by the front cavity in the second model. The high damping associated with the surface losses in the narrow tube and the radiation resistance affect B_3 of (A) and B_2 and B_4 of (B).

The two models do not differ in terms of B_1. Theoretically it would be possible to choose the correct l_1, l_2, A_1, A_2 of the two-tube model from a specification of F_1, F_2, F_3 and either B_2 or B_3 or the ratio B_2/B_3 or B_4 or some combination of B_4 and other bandwidths, e.g. $(B_2+B_4)/B_3$. In a real speech case the situation might be different if the glottal losses are large and execute high damping of the back tube resonances.

In practice it may take a ventriloquist to produce something similar to these two models. Possibly the one with a shorter back tube would fit into the vocal tract anatomy of a very small child, as suggested in the previous section.

In conclusion - to improve techniques for inferring vocal tract characteristics from speech wave data we need a better insight in vocal tract anatomy, area function constraints, and a continued experience of confronting models with reality - a balanced mixture of academic sophistications and pragmatic modeling.

References

Atal, B.S., J.J. Chang, M.V. Mathews, and J.W. Tukey (1978): "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", JASA 63, 1535-1555.

Fant, G. (1960): Acoustic theory of speech production, The Hague: Mouton (2nd edition 1970).

Fant, G. (1965): "Formants and cavities", Proc.Phon.5, 120-141, Basel: Karger.

Fant, G. (1966): "A note on vocal tract size factors and non-uniform F-pattern scalings", STL-QPSR 4, 22-30.

Fant, G. (1972): "Vocal tract wall effects, losses, and resonance bandwidths", STL-QPSR 2-3, 28-52.

Fant, G. (1973): Speech sounds and features, Cambridge, Mass.: MIT Press.

Fant, G. (1975a): "Non-uniform vowel normalization", STL-QPSR 2-3, 1-19.

Fant, G. (1975b): "Vocal-tract area and length perturbations", STL-QPSR 4, 1-14.

Fant, G. (1976): "Vocal tract energy functions and non-uniform scaling", J.Acoust.Soc.Japan 11, 1-18.

Fant, G. (1979): "Glottal source and excitation analysis", STL-QPSR 1, 85-107.

Fant, G. and S. Pauli (1975): "Spatial characteristics of vocal tract resonance modes", in Proc. Speech Comm. Sem. 74: Speech Communication, Vol. 2, G. Fant (ed.), 121-132, Stockholm: Almqvist and Wiksell.

Fant, G., K. Ishizaka, J. Lindqvist, and J. Sundberg (1972): "Subglottal formants", STL-QPSR 1, 1-12.

Fant, G., L. Nord, and P. Branderud (1976): "A note on the vocal tract wall impedance", STL-QPSR 4, 13-20.

Fant, G. and J. Liljencrants (1979): "Perception of vowels with truncated intraperiod decay envelopes", STL-QPSR 1, 79-84.

Flanagan, J.L. (1965): Speech analysis synthesis and perception, Berlin: Springer (2nd expanded ed. 1972).

Flanagan, J.L., K. Ishizaka, and K. Shipley (1975): "Synthesis of speech from a dynamic model of the vocal cords and vocal tract", Bell System Techn. J. 54, 485-506.

Fujimura, O. and J. Lindqvist (1971): "Sweep-tone measurements of vocal-tract characteristics", JASA 49, 541-558.

Goldstein, U. (1979): "Modeling children's vocal tracts", JASA 65, S25(A).

Guérin, B., M. Mrayati, and R. Carré (1975): "A voice source taking into account of coupling with the supraglottal cavities", Rep. from Lab. de la Communication Parlée, ENSERG, Grenoble.

Heinz, J.M. (1967): "Perturbation functions for the determination of vocal-tract area functions from vocal-tract eigenvalues", STL-QPSR 1, 1-14.

Ishizaka, K., J.C. French, and J.L. Flanagan (1975): "Direct determination of vocal tract wall impedance", IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP-23, 370-373.

Ishizaki, S. (1975): "Analysis of speech based on stochastic process model", Bull. Electrotechn. Lab. 39, 881-902.

Jospa, P. (1975): "Effets de la dynamique du conduit vocal sur les modes de résonances", Rep. de l'institut de phonétique, Université Libre de Bruxelles, 51-74.

Ladefoged, P., R. Harshman, L. Goldstein, and L. Rice (1978): "Generating vocal tract shapes from formant frequencies", JASA 64, 1027-1035.

Lindblom, B. and J. Sundberg (1969): "A quantitative model of vowel production and the distinctive features of Swedish vowels", STL-QPSR 1, 14-32.

Lindqvist, J. and J. Sundberg (1972): "Acoustic properties of the nasal tract", STL-QPSR 1, 13-17.

Mrayati, M. and B. Guérin (1976): "Etude des caractéristiques acoustiques des voyelles orales françaises par simulation du conduit vocal avec pertes", Revue d'Acoustique 36, 18-32.

Mrayati, M., B. Guérin, and L.J. Boë (1976): "Etude de l'impédance du conduit vocal - Couplage source-conduit vocal", Acustica 35, 330-340.

Nordström, P.-E. (1975): "Attempts to simulate female and infant vocal tracts from male area functions", STL-QPSR 2-3, 20-33.

Öhman, S.E.G. and S. Zetterlund (1975): "On symmetry in the vocal tract", in Proc. Speech Comm. Sem. 74: Speech Communication, Vol. 2, G. Fant (ed.), 133-138, Stockholm: Almqvist and Wiksell.

Schroeder, M.R. (1967): "Determination of the geometry of the human vocal tract by acoustic measurements", JASA 41, 1002-1010.

Sidell, R.S. and J.J. Fredberg (1978): "Noninvasive inference of airway network geometry from broadband long reflection data", J. of Biomedical Eng. 100, 131-138.

Sondhi, M.M. and B. Gopinath (1971): "Determination of vocal tract shape from impulse response at the lips", JASA 49, 1867-1873.

Stevens, K.N. (1971): "Airflow and turbulence noise for fricative and stop consonants, static considerations", JASA 50, 1180-1192.

Stevens, K.N. and A.S. House (1955): "Development of a quantitative description of vowel articulation", JASA 27, 484-493.

Wakita, H. (1973): "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms", IEEE Trans. Audio and Electroacoustics, AU-21, 417-427.

Wakita, H. (1979): "Estimation of vocal tract shapes from acoustical analysis of the speech wave: the state of the art", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-27, 281-285.

Wakita, H. and G. Fant (1978): "Toward a better vocal tract model", STL-QPSR 1, 9-29.

DISCUSSION

Hisashi Wakita, Raymond Descout and Peter Ladefoged opened the discussion.

Hisashi Wakita: In determining the interrelationship between speech articulation and acoustics, we are particularly interested in the inverse problem, i.e. the estimate of vocal tract shapes from the acoustic waveform. There are various uncertain factors in deriving vocal tract area functions from the waveform, but it is an attractive method, because it is both the safest and easiest. (The problem with recent articulatory models for vocal tract shaping is that we do not yet know the exact parameters that control vocal tract shapes, in terms of articulators, and we do not have sufficient methods to obtain the data.) One of the most promising methods is the linear prediction (LPC) method, to estimate area functions from acoustic data. We do not know to what extent we can describe the details of the vocal tract shape, but by combining the LPC method with physiological data, we hope to improve this method.

One problem is the non-uniqueness, i.e. we can generate an infinite number of shapes having exactly the same frequency spectrum within a limited frequency band. To solve the uniqueness problem we have to impose constraints, physiologically determined constraints, or constraints determined by the higher harmonic structure. So far, the LPC method has been using formant frequencies and bandwidths, and in fact the final area function is sometimes quite sensitive to bandwidth. But we would like to get rid of bandwidth in the calculations: From the first three formant frequencies we can obtain the midsagittal view of the vocal tract, like in the Peter Ladefoged model, and to get at the unique shape of this midsagittal area function we may employ physiological constraints.

Another problem with LPC analysis is the vocal tract excitation and the losses, both within the vocal tract and at its boundaries, and these problems have to be solved in order to get more accurate vocal tract shapes. In fact, with the LPC method we can detect the closed glottis portion, where the interaction between sub- and supraglottal cavities is minimized, which makes for more accurate area functions. A further draw-back of LPC is

that we have to start from the very simple assumptions of a simple loss at the glottis and a lossless acoustic tube. On the other hand, you can make a production model as complex as you wish, - you can add any realistic losses along the vocal tract or at the glottis that you like, but as analysis model there is a strong limitation in incorporating losses and other factors. So at this moment, the imminent problem is how to attack the loss problems and the source uncertainties.

Raymond Descout: Very little original data has accumulated on area functions, because collecting it is difficult, from a technical point of view. On the other hand, deriving vocal tract area functions from acoustic data has some disadvantages: with LPC techniques we only get pseudo area functions, and with acoustic measurements, which I previously worked on, there is a great problem in dynamic measurements, especially. Further, interest has largely centered on the midsagittal view of the tract, but we need information about the frontal view as well, which may be obtained with the new techniques of computerized tomography. We need this information in order to turn the midsagittal view into a three-dimensional area function, and to determine the shape factors that are necessary for the introduction of losses in our models.

All the articulatory models proposed are based upon vowel configurations, and when we try to make dynamic simulations on the articulatory model, everything that we do not know about the consonants is put into a special coarticulation and transition rule. We need more information on the consonants.

The acoustic model of the vocal tract is derived from the propagation equations, based on assumptions of symmetrical, equal length sections, - but to do an inverse transform you really need a very appropriate model which includes the shape factors that are necessary for the loss calculations, because the mathematical technique involved in the transformation is stupid in the sense that the result will be adjusted according to mathematical criteria, but this may not result in a realistic vocal tract. Therefore, I think that doing inverse vocal tract transforms is premature: we must work first of all on the proposition of the best production model, including shape factors and losses, before trying to do inverse vocal tract transforms.

Due to the progress made in articulatory modelling and to the limitations of LPC-techniques, we have witnessed a come-back of studies on vocal tract and vocal source simulations. To refine the articulatory model, we need further physiological data.

In conclusion: I do not think that LPC will give us a better understanding of speech production (it is, however, excellent for synthesis purposes). We need more studies on the relationship between articulatory parameters / area functions / vocal tract shapes.

Peter Ladefoged: Gunnar Fant showed us many years ago that what is important in characterizing speech are the first three formant frequencies, and you can even get a great deal of a speaker's personal quality with just three formant frequencies. But with the inverse transform, to get as far as eight tubes (which is only a coarse model of the vocal tract), you need at least four formant frequencies and their bandwidths, and with eighteen tubes you need nine formant frequencies and bandwidths, etc. Now something is wrong here: any phonetician can draw, more or less accurately, the midsagittal view of a given speaker's vowels, and we ought to be able to develop an algorithm that will go from the acoustics to the tract shape. There are of course problems - we do not actually observe the tract shape, only the midsagittal dimensions, and there are only very limited sets of data that tell us how to derive the tract shape from the sagittal dimension.

The work of Lindblom and others has shown that you can produce an [i:] with your jaw in a more or less open position, i.e. one has the ability to control tract shapes using different articulatory procedures, and it is of great interest to us to know how we exert that control and less interesting what the muscles do. Eventually, we have got to be able to go from acoustic structures, finding out what the tract shape is, and then deducing from that what the underlying control signals must have been.

Gunnar Fant: I agree with the main points of the discussants. Inverse transforms cannot make up for our great lack of physiological reference data.

My suggestions for improving inverse transform techniques

in part supported by the previous discussions are: (1) we should model the vocal tract in terms of lossy transmission line sections instead of the simplified LPC model, (2) we should not expect to generate a larger number of independent production parameters than we have independent and well specified speech wave descriptors relating to the vocal tract transfer function. Overspecified area functions are necessarily non-unique, whereas a balanced specification can be, but need not be, unique. With proper model and parameter constraints, a 32-section area function model may be generated from a set of 3-6 articulatory parameters and controlled by the same number of acoustic parameters. It remains to be seen if we can extract more than four independent acoustic parameters. (3) The vocal tract total length should be derivable from one extra independent acoustic parameter.

Our discussion concerning bandwidths is still rather academic and we appear to share a doubt concerning the specificational value of bandwidths. Theoretically the set F_1 F_2 B_1 B_2 could suffice to specify a three-parameter model extended with a fourth parameter, e.g. the total length. This might hold for a resonator model only but not for a true vocal tract with less predictable bandwidth sources and the limited accuracy in bandwidth measurements. A more efficient set of acoustic parameters would be F_1 F_2 F_3 and B_3. From my Fig. 16 illustrating bandwidths of Swedish vowels it is seen that B_3 is a good correlate of degree of lip opening and also mouth opening. However, vowel bandwidths including B_3 are to a high degree predictable from formant frequencies. The role of bandwidths in an LPC model is not the same as that of a true vocal tract model. This is an important distinction. The LPC bandwidths, e.g. B_3, may come out quite different from those of real speech or from simulations by an improved model. The bandwidths we need for the inverse LPC based transforms are the bandwidths of a production model which has losses at the glottis only and locks the cavity wall shunt. From the true formant frequencies and bandwidths we thus have to make a best guess of what bandwidths the LPC model would generate. This is in the line of the recent work of Hisashi Wakita (1979).

Kenneth Stevens: With regard to what a male speaker does in order to compensate relative to the [u:] of a female: if we define narrow vowels as having so narrow a constriction that turbulence is just not generated, is it conceivable then that males, who generate a greater air flow than women, cannot round the vowels as much as can women, and therefore the formants are not lower than those of women?

Gunnar Fant: It could be, but in Swedish the vowel [u:] as well as [i:], [y:], and [ʉ:] are generally produced, by males and females alike, with a diphthongal glide passing through a relatively constricted phase in which some turbulence may be generated. I would rather expect different male and female articulations to be aimed at some criterion of perceptual invariance of which we do not know too much yet.

Antti Sovijärvi asked Gunnar Fant what his concept is about nasalized vowels.

Gunnar Fant: An essential characteristic of nasalization independent of the specific resonances added is the reduced F1 amplitude which is especially apparent in an oscillographic analysis. What appears to be a sub-F1 nasal formant is often a voice source feature which is relatively re-inforced because of the F1 reduction.

Hisashi Wakita: As long as the calculations are based on the first few formant frequencies, the problems in inverse transformation are rather equivalent with different methods. To uniquely determine a six tube vocal tract shape, LPC uses the first three bandwidths. If you want a smooth area function, you have to specify one of the higher frequency characteristics, and to do that you have to impose some kind of constraint, which is what Dr. Ladefoged does. And whatever the method, if you do not want to use bandwidth, you have to use some other kind of information to uniquely determine the spectra, and any information will do as long as you are able to reconstruct the original spectrum with its original bandwidths - so bandwidth is in fact a very important parameter.

Gunnar Fant: It would be interesting to see how far you would get if you started out with F1, F2, and F3 and then predicted B1, B2, and B3 from the formulas that I have.

Peter Ladefoged: I have tried using Hisashi Wakita's formulae with Gunnar Fant's type of predicted bandwidths (and other bandwidths from the literature), and it did not work, - I got absolutely impossible vocal tract shapes. Regarding Atal's vocal tract shapes that produce identical formant frequencies: some of them are quite impossible, the tongue just cannot produce some of those shapes.

John Holmes: I wish to emphasize the difficulty of mathematically deriving the vocal tract from the speech waveform, because we know too little about the glottal source. Gunnar Fant emphasized that the closed glottis portion is better suited than the open glottis portion to work out the supraglottal characteristics, but (as can be seen on the Farnsworth vocal chord movie of about 1940 and from Tom Baer's work), even when the vocal chords are closed there is sufficient ripple and surface movement for there to be an effective volume velocity input into the vocal tract, which means that your resultant waveform is never a force-free response, - and this is one of the things that makes bandwidths so difficult to estimate, because it is quite possible that ripple in vocal chord surface could actually be causing the formant amplitude to be still building up even, in exceptional cases, during the closed glottis period. I think this supports the view that we have to work from much more basic information and use articulatory constraints rather than to derive vocal tracts by purely mathematical techniques from some artificial and unrealistic production model.

Gunnar Fant: I can only agree with your statements. It is necessary to learn more about the human voice source in order to improve our methods of inverse transforms.

Osamu Fujimura: We can obtain cross-sectional vocal tract shapes with the regular computerized tomography, but only at great costs, because the X-ray dosage is tremendously high, a requirement of brain diagnoses that demand a very good density solution.

But I think the machine can be adjusted and the X-ray dosage reduced for our purposes, where we are really only interested in the distinction between matter and air.

Mohan Sondhi at the Bell Laboratories has proposed an acoustic impedance measurement using an impulse-like excitation at the lips, which can give us complete information about the area function of the vocal tract, because we obtain two sets of infinite series, i.e. the poles and the zeroes of the impedance function that together uniquely determine the vocal tract shape, without having to assume or measure losses. I think that there is one major difficulty with this technique: the subject articulates silently, i.e. he has no auditory feed-back, and we cannot be sure about the actual gestures. That problem can be overcome if we simultaneously monitor the vocal tract with e.g. the X-ray microbeam method.

Gunnar Fant: The micro-beam system will certainly provide us with excellent data about speech articulation, but will it provide us with all the details that we want about the vocal tract, like the exact dimensions of the pharynx and larynx cavities?

Osamu Fujimura: We can obtain data on cross-sectional shapes, because we can place pellets also outside the midsagittal plane, - the only constraint being that we cannot use too many pellets at the same time, which will increase the X-ray dosage, but it is not easy to place pellets on the pharyngeal walls, which is a limitation of the method. However, we have a new stereo-fiberscope which can be used for three-dimensional optical observations of the pharynx, and I hope in the future to be able to develop a technique that will supplement the X-ray technique with this kind of optical information.

Raymond Descout: I am presently working with a prototype CT (computerized tomography) scanner, which scans in five seconds, and we are trying to lower the X-ray dosage to ten percent the normal dosage, because all we need is to see the difference between air and flesh. There is still a problem with the CT technique, though, and that is determining exactly the position of the slice relative to the skin and the rest of the person.

MODERN METHODS OF INVESTIGATION IN SPEECH PRODUCTION

Osamu Fujimura, Bell Laboratories, Murray Hill, New Jersey 07974

Chairperson: Celia Scully

1. Descriptive Theory and Modeling of Speech Production

The process of speech production involves many aspects which may be treated by different disciplines of science. As much as we deal with speech as signals representing linguistic codes, it is clear that we need to have a descriptive framework of the linguistic message, so that we can relate the observed physical phenomena to the units that are used in the codes. Both segmental and suprasegmental specifications have to be given, as well as appropriate indications of surface syntactic (and semantic) information.

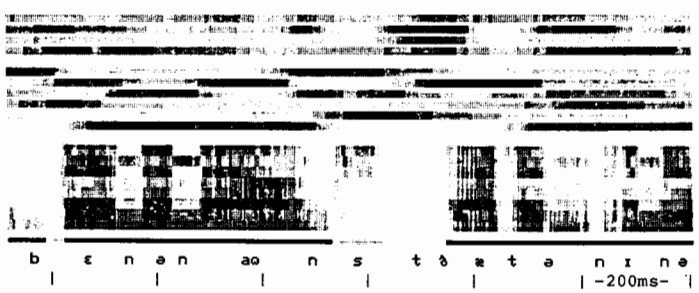
In addition to the lexically distinct accentual patterns and different intonational patterns for phrase structures, modulations of voice pitch and duration may be extensively used in conversational speech reflecting, e.g., focus, emphatic contrast, contextual and statistical predictabilities of the word, etc. Since speech phenomena always involve paralinguistic factors, such as the speaker's emotional state and idiosyncrasy, a way of describing those is also needed; or at least we must have a clear idea about what relevant factors have to be kept constant to make the comparison of different linguistic units meaningful. These considerations become more and more important, as we make progress in speech research. There are some emerging efforts in this direction, both in theory and experiment. The metric theory (Liberman and Prince 1977) for description of stress and intonation patterns of English constitutes a good example of such theoretical progress in this area, and a pitch contour synthesis-by-rule experiment based on this theory (Pierrehumbert 1979) suggests rapid progress in this field.

The notion of segments is also being revisited in connection with the significance of larger segmental units. The basic idea is to concatenate segmental units, whether phonemes, syllables, phonological words or phrases, to form larger units, and give suprasegmental modulations as patterns assigned to the larger units. Experiments in synthesis by rule attempt to evaluate models of this process. The notion of temporal modulation can be clarified only by referring to a well-defined model of speech dynamics that im-

plements an abstract specification of concatenated strings of units. Such a phonetic realization process would be characterized by different dynamic (i.e. temporal) characteristics for individual articulators, and the realized phonetic events corresponding to the so-called (phoneme size) "segments" are in general not in synchrony. Therefore, discontinuities observed in acoustic signals, such as the voice onset, stop release, etc., may not reveal some of the important aspects of the temporal characteristics of speech.

Gunnar Fant (1962) described a fine subsegmentation of acoustic signals based on their apparent discontinuities and interpreted such spectrographic representations of speech in terms of overlapping acoustic properties, roughly similar to, but crucially different from, the linguistic distinctive features.

FIGURE 1

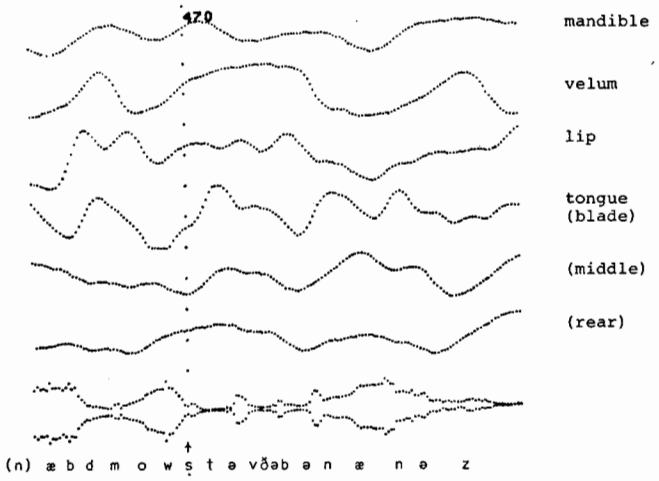


A combined articulatory-acoustic representation of part of a sentence 'Ben announced that an innocent-seeming infant had nimbly nabbed most of the bananas', uttered by a male native speaker of American English (Fresno, California). The upper part pertains to pellet positions, as obtained by the computer-controlled x-ray microbeam system, and the lower part a simplified (8 frequency-band) spectrographic pattern. In the lowest horizontal line black, gray, and white represent, respectively, voiced, voiceless, and silent states of the speech signal, and the phonetic symbols underneath are selected and placed automatically based on the articulatory information as well as the voicing state of the sound. The articulatory gesture is represented by the topmost 4 stripes for front (dark)/back (light) movements of the pellets placed on (from top) the lower lip, the blade, mid and rear portions of the tongue, and below these by the 6 stripes for up (dark) - down (light) movement of (from top) the lower lip, the mandible, the three parts of the tongue, and the velum (dark for low) (see Nelson [1979], Miller and Fujimura [1979]).

In order to account for the full information contained in speech signals and its human perception, one has to go well beyond this basic sketch. The spectral modulation of the speech signal is in one aspect discontinuous and in the other continuous. This dual nature of speech may be seen most obviously when we compare a gross spectrographic representation with an articulatory representation (see Figure 1) (Miller and Fujimura 1979). This qualitative difference betweeen articulatory movement and its consequent acoustic temporal pattern stems from the inherent non-linearity between the two levels of speech representation.

Recent studies are revealing interesting details of articulatory processes in relation to the phonological structures of the message. It is being shown that a simple model of concate-

FIGURE 2



Time functions representing vertical movements of the 6 pellets (the same material as in Fig. 1). The lowest trace depicts the speech waveform envelope. The arrow in line with a vertical array of dots is placed at the beginning of the voiceless segment for /st/.

nating phoneme-size units into larger phonological units, taking care of "coarticulation" phenomena by smoothing the movements, simply does not work. This is so particularly because within each syllable (or more exactly syllable core, see Fujimura and Lovins (1978), Fujimura (1979b)) there is something much more ad hoc about the temporal structure of phonetic events as syllabic ingredients. Such ad hoc characteristics are largely dependent on the language (and dialect) and therefore cannot be specified by a universal phonetic principle. By examining articulatory processes for relevant organs in movement, allowing for different dynamic characteristics and freedom of asynchrony in motor control for different articulatory (or phonatory) dimensions, we can obtain some insight into the nature of the temporal organization of phonetic events (Fujimura, forthcoming-a). Even inversions of temporal relations of peak activities for individual articulatory gestures are observed, from a phoneme string point of view. For example, as shown in Fig. 2, the syllable /mowst/ in a sentence utterance shows that the labial constriction for the glide /w/ manifests its peak activity during the voiceless period for /st/ toward the articulatory closure of /t/. A general principle governing phonetic structures of syllables (for the language) guarantees this looseness of temporal ordering within the syllable core to be irrelevant for phonological identification of this form (see Fujimura and Lovins (ibid)).

A useful descriptive framework thus seems to be one based on individual articulatory events related to elementary (functional) features of the syllable core. Basic notions, such as concatenation, coarticulation, assimilation and dissimilation have to be revisited quantitatively in light of such a descriptive model. It is time for us to produce experimental evidence for or against specific intuitive predictions. The scope of such experimental work is now being drastically expanded, thanks to newly available tools. It must be emphasized, however, that any of the available techniques for physical measurement, even in the future, is not likely to provide us with a complete picture of the physiologic/physical phenomena of speech production by itself. In order to interpret the results of measurements at different levels and relate them to each other, which is the task given to speech scientists for understanding the speech production process, we need to devise some new tools. Computational models of the natural speech production apparatus are being studied as such tools. For example, a three-dimensional static model of the tongue has been constructed using the finite element method (Kiritani et al. 1976) and is being used for studies of control characteristics of vowels (Fujimura and Kakita 1979).

A quantitative study of the gesture for the vowel [i], based on the tongue model, has suggested that the contraction of the posterior portion of the genioglossus muscle alone can give rise to a reasonable shape of the tongue and a consequent formant pattern for this vowel, but a slight deviation from the correct magnitude of contraction would cause quick deviation from the acceptable phonetic value. On the other hand, if we use a set of muscle components, in conformity with available electromyographic findings, we find such sensitivity to the degree of contraction is eliminated and the resultant phonetic quality becomes very stable and easy to achieve with a wide latitude of physiologic control. This points to the question of the quantal nature of speech as proposed by K. N. Stevens (1972), and also to the significance of feedback in different situations of speech production, including normal and artificial (such as the bite block) circumstances. With respect to the quantal nature, it seems that the crucial issue is the choice of the input level, at which the change of the controlled quantity in question is compared with that at the output, i.e. the acoustic characteristics such as formant patterns. The midsagittal tongue contour or a parametrically represented area function does not seem to be the correct input to the system for this specific discussion. The three-dimensional structure of the tongue combined with its volume incompressibility seems to play an essential role in characterizing the nonlinearity of the input-output mapping. Also, if, as our tongue model study seems to suggest, what is important in achieving a phonetic goal of articulatory gesture is selecting the pertinent set of muscles (with a certain balance of relative activities) rather than the exact magnitude of muscular contraction (excessive contraction resulting only in more or less unaffected physical consequences) the observed robustness of articulation under affected conditions seems more readily explicable than we had thought before. Gross orosensory feedback information also seems to play an important role in this connection (Perkell 1977).

Within the hierarchy of the natural process of speech production, the higher the level, the less applicable direct physical measurements are. Recent efforts by psychologists (see e.g. Sternberg et al., (1978)) are focused on temporal aspects of motor control, in the attempt to infer basic mechanisms of cortical programming and its execution. Studies of highly skilled performances in nonspeech areas seem to point to the understanding that in routine human actions the temporal course of a physical state takes a fixed preprogrammed pattern. In speech, articulatory events are decomposable into elementary gestures, such as lip movements for bilabial stops and velum raising for nasal-to-non-nasal transitions. Recent articulatory measurements indicate relatively constant speeds of such movements in a wide range of conditions when influences of certain separable factors are excluded (see the co-report on speech production by Sawashima (vol. I, p. 49-56)).

It has been argued (MacNeilage 1970) that the notion of invariant gestures for phonetic units is untenable in consideration of the high number of different contextual conditions. Such estimates, however, customarily depend on phoneme-size phonetic units as the basis of assuming targets. Based on an analysis that syllables are separable into cores and phonetic affixes, and each core into relatively constant dynamic patterns of initial and final demisyllables (the latter including the central portion of the syllable), we can actually construct for English a complete inventory of phonetic (concatenative) segmental units that contains less than 1,000 items for virtually all possible English phonetic forms (Lovins et al., 1979). Assuming that each inventory item is given phonetic indexes (syllable features) representing articulatory gestures, and also temporal parameters that are sensitive to nonsegmental conditions such as stress/accent, speed of utterance, etc., it does not seem implausible that the human brain can store all necessary phonetic patterns in the given language. An experimental evaluation of this new view is being attempted by synthesis-by-rule experiments using a demisyllabic inventory. A concrete model of acoustic realization of syllable features is being studied by Mattingly (1977). The psychological reality of the core-affix decomposition as well as the syllable itself is still to be examined.

2. Physiological Studies - Muscle Controls

The study of the physiology of speech production has seen remarkable progress in the past decade, even though there are still many unsolved basic questions. One general question is which muscle plays the principal role of implementing motor commands for a given phonetic gesture, viz. an elementary articulatory event. Electromyographic studies have revealed, for example, that the glottal abduction reflecting the devoicing gesture is related to the activity of the posterior cricoarytenoid muscles, whereas glottal adduction is achieved by several different muscles, including the interarytenoids, in varied ways depending on linguistic (and paralinguistic) functions (Hirose and Gay 1972; Hirose et al. 1978).

Hirano recently studied the anatomy and physiology of the vocal cords using various advanced techniques such as electron microscopy, histochemistry, electromyography, electric nerve stimulation, high speed motion picture, mechanical measurements, applied to both human and animal larynges (Hirano 1977). He arrived at an approximation of the complex anatomical structure by two (or three) loosely coupled parts, viz. cover and body. The cover seems to be responsible for the major part of the vibratory movement, showing large three-dimensional excursions, whereas the body contains the so-called vocalis muscle and participates in active parametric control of the vibrating system (see also Fujimura (1979a)). Baer (1975) has contributed a detailed study of excised canine larynges, and Titze and Talkin (1979) are contributing a new computerized model of the vocal cord vibration process.

Pitch control is an important topic from both lexical distinction and sentence-intonation points of view. The physiologic mechanism is not completely understood, but much is known now about the function of the cricothyroid muscle in relation to the voice fundamental frequency. There are cases where the voice fundamental frequency does not reflect the phonological accentual pattern because of the interaction between the consonantal control of voicing/tenseness and the vocal fold vibration frequency, but the electromyographic signal of the cricothyroid does (see Fujimura (forthcoming-b)).

Lingual muscles are difficult to study even with the best available electromyographic techniques because of the complex

interdigitation of a number of muscles forming the main body of the tongue. Nevertheless, the rather limited information obtained by EMG measurements are indispensable in inferring muscular functions relative to specific phonetic gestures.

Controlled interference by such techniques as anesthesia and bite block, has been experimentally induced, in order to evaluate the roles of feedback loops in speech production (Lindblom et al. 1977). In real utterance situations, mandible height is not necessarily correlated with tongue height either positively or negatively. For example, for the American English vowels /eɪ/ and /ɛ/ in sentence utterances, we have found in our X-ray microbeam data that a tongue height measure does distinguish occurrences of the two vowels very clearly, but that mandible height can be either lower or higher for one vowel than the other. Mandible height seems to reflect the stress status of the vowel, serving a function that is partially independent of the vowel height specification.

3. Physical States of Organs

Neural control of the larynx is parametric in the sense that gross average states of the larynx rather than details of vibratory changes of the peripheral shapes of the vocal cords are adjusted. For this reason, if we measure the laryngeal state during an utterance, the measurement may be taken at a relatively slow sampling rate such as 50 samples/second and averaged over a period like 20 msec. The fiberscopic technique developed at the University of Tokyo is appropriate for this purpose (Sawashima and Hirose 1968).

There have been successful studies of segmental control, such as manners of consonantal articulations in different languages (see for a review, Fujimura (1979a)). Here again, there are cases where the acoustic signal cannot answer a question about control. The laryngeal maneuver for pitch control seems related to vertical movements of the larynx as well as other gross appearances of the glottal area, and this may give us an opportunity to learn about pitch control even for devoiced syllables. A recent improvement of the fiberscope has made it possible to record two images side by side on the film stereoscopically, so we can measure the distance between the objective lens and the object (Fujimura et al. 1979). For many phonetic studies on qualitative states of the glottis, on the other hand, electric resistance measurements are being used as a readily applicable tool (Fourcin 1977, Frøkjær-Jensen 1968). Characteristics of voice source signals have gained renewed interest. Gunnar Fant (this volume, p. 79-108) is contributing a new insight about the interaction between the source and the vocal tract by closely examining speech waveforms. Flanagan et al. (1975) used their two-mass model of the vocal cords for simulating turbulence generation in the coupled source-vocal tract system.

The lips are obviously the easiest object to measure among different articulators, particularly with the use of a powerful stroboscopic technique (Fujimura 1961). A modern computerized system for measurement of the lips and mandible positions as well as linguapalatal contact is now available at the University of Alabama (McCutcheon et al. 1977). A servomechanistic technique can be used for a more general analysis of the natural articulatory systems such as the mandible and the lips. Such a measurement system has been implemented at the University of Wisconsin, Madison, and the control mechanisms of the lips are being studied assuming a linear system with feedback loops (Muller and Abbs 1979). The frequency response of such looped systems seems to allow actively controlled movements of visco-elastic systems via brainstem feedback for the majority of speech events. It should be emphasized, however, that the peripheral parts of articulators do not necessarily move together with the neurally controlled body of the same organ, and it is the former that determines acoustic consequences.

Dynamic characteristics of articulators in speech have been a vital issue in speech research. Several interesting proposals have been made about the basic principle of articulatory gestures trying to relate abstract and discrete phonological codes to the temporal structures of continuous speech phenomena (see Kent and Minifie (1977) for a review). Information on actual movements of the principal organs, in particular the tongue, is badly needed for such a study. Relatively large amounts of data obtained from the same subject are necessary to cope with an inherent variability of speech production phenomena. Collection of comparable data from many subjects, wherever possible, is another necessity for understanding the other aspect of human variability.

There are several methods that have been proposed and tested for observing tongue movements. Dynamic palatography (Fujimura et al. 1973b) represented an early attempt to computerize tongue observation for acquisition and processing of large amounts of data. It is also being applied to training of children in speech and hearing clinics in Japan. Other more recently proposed techniques include optical distance measurement between selected points on the palate and the nearest tongue surface. Magnetic (Sonoda 1977) as well as ultrasonic (Minifie et al. 1971) measurements also have been proposed.

The most direct and informative method of observing tongue movement is the use of X-rays for lateral views of the tongue. There used to be two factors that made radiographic measurements impractical for obtaining a large quantity of speech data. One is the radiological disturbance given to the subject. For this reason the exposure had to be limited usually to one or two minutes total per subject. The tedious and inefficient frame-by-frame analysis of the photographic images constituted another problem. The computer-controlled X-ray microbeam system was devised precisely to overcome these difficulties (Fujimura et al. 1973a). A full-scale system is now in operation at the University of Tokyo (Kiritani et al. 1975), and is producing useful results.

Several metal pellets are placed on selected points on the tongue and other articulators, usually but not necessarily in the midsagittal plane. A computer directs a thin X-ray beam to search around a predicted position, for each pellet, based on its past position and movement, verifies the current position, and repeats the procedure to look for the next pellet. By the combination of high sensitivity of the X-ray detector and an efficient use of the given total dosage for determining pellet positions, without exposing any unnecessary portions of the body for the specific purpose, the total radiographic exposure is incomparably smaller than that which would be used by film recording with an image-intensifier. The pellet position at each sample time, typically every 10 ms or less for 6-8 pellets, is digitally stored in the computer memory in real time. The experimenter, and the subject if desirable, can monitor the detected pellet movements. Powerful computer programs have been designed and implemented at Bell Laboratories in order to give the experimenter an efficient tool for interactive data analysis. Figure 1 represents one of the results, including an automatic annotation of the speech material with phonetic symbols (Nelson 1979).

An independent estimation of area functions by acoustic input impedance measurement has been proposed (Sondhi and Gopinath 1972). There is a nontrivial mapping process between the acoustically effective area function and the state of the speech organs (Mermelstein 1973). On the other hand, the so-called pseudo-area function that is conveniently derived by the well-established linear-prediction coding scheme (LPC) is not a true representation of the vocal tract characteristics proper (see Fant, p. 79-108, and Wakita, p. 151-172 (this volume)). Therefore, it is very desirable to have such independent measurement of the true area function, particularly if a simultaneous X-ray observation can be made for direct comparison of tongue shape (pellet positions) and the effective area function. The use of the recently developed CAT technique is also being attempted for static gestures.

4. Statistical Processing of Production Data

The availability of a large amount of production data encourages researchers to use advanced techniques of statistical processing of data such as multidimensional analysis (INDSCAL (Carroll and Chang, 1970) or PARAFAC (Harshman et al. 1974)), as well as principal component analyses. Through purely statistical processes, constituent (static) gesture components have been derived from both hand-traced midsagittal contours of the tongue of many speakers (Ladefoged 1977) and automatically tracked pellet position data for each of a few speakers (Kiritani and Imagawa 1976). These inductive methods give us purely phenomenologically derived "phonetic coordinates" for describing articulatory characteristics of a class of phonetic units, which is defined by the particular choice of the speech material used for this data processing. It is an intriguing question to ask if we can have a universal descriptive framework that explains the relations between different aspects of categorization of phonetic units (see Ladefoged's co-report on speech production (vol. I, p. 41-47)).

The use of multiple regression technique (both linear and nonlinear) must be mentioned in connection with the inverse mapping from acoustic characteristics to articulatory conditions. In addition to the more traditional method of analysis-by-synthesis,

which also is being used extensively (Fujisaki 1977), such new computational means seem to promise a new trend of research. Multiple regression techniques have been used for interpreting both durational parameters (Liberman 1978), and articulatory data (Nakajima 1977, Shirai and Honda 1977). The former used automatic processing of reiterant speech signals (Liberman and Streeter 1978), having the subjects mimic a sentence by a repetition of the same syllable, such as [ma], and attempted a best match between model-predicted and measured syllable durations by adjusting relative contributions of different phonologic and syntactic factors. The latter, using nonlinear regression, assumes a simple dynamic model of the physical movements of the articulators to determine the parameters that characterize such a physical system.

5. Concluding Remarks

When we define a domain of problems, such as normal speech, speech of a particular speaker, vowels as opposed to consonants, phonology as opposed to syntax, etc., we always need some understanding of the problems surrounding that domain. By knowing what happens just outside the boundary of the domain of immediate interest, in accordance with the principle of continuity, we always gain better insight as to how to delimit the domain. Thus, for example, speech pathology is another intriguing area of phonetic research. Needless to say, we would like to learn how people perceive speech, in order to investigate how people speak, because the real-life speech behavior is always a continuous mixture of production and perception.

References

Baer, T. (1975): Investigations of phonation using excised larynxes, PH.D. dissertation, M.I.T.

Carroll, J.D. and J. Chang (1970): "Analysis of individual differences in multidimentional scaling via an N-way generalization of 'Eckart-Young' decomposition", Psychometrika 35, 283-319.

Fant, G. (1962): "Descriptive analysis of the acoustic aspects of speech", Logos 5, 3-17.

Flanagan, J.L., K. Ishizaka, and K.L. Shipley (1975): "Synthesis of speech from a dynamic model of the vocal cords and vocal tract", Bell Syst. Tech. J. 54, 485-506.

Fourcin, A.J. and E. Abberton (1977): "Laryngograph studies of vocal-fold vibration", Phonetica 34, 313-315.

Frøkjær-Jensen, B. (1968): "Comparison between a Fabre glottograph and a photo-electric glottograph", Annual Report of the Institute of Phonetics, University of Copenhagen 3, 9-16.

Fujimura, O. (1961): "Bilabial stop and nasal consonants: A motion picture study and its acoustical implications", JSHR 4, 233-247.

Fujimura, O., S. Kiritani, and H. Ishida (1973a): "Computer controlled radiography for observation of movements of articulatory and other human organs", Comput. Biol. Med. 3, 371-384.

Fujimura, O., I.F. Tatsumi, and R. Kagaya (1973b): "Computational processing of palatographic patterns", JPh 1, 47-54.

Fujimura, O. and J. Lovins (1978): "Syllables as concatenative phonetic units", in Syllables and segments, A. Bell and J.B. Hooper (eds.), 107-120.

Fujimura, O. (1979a): "Physiological functions of the larynx in phonetic control", in Current issues in the phonetic sciences (Proc. of the IPS-77 Congress, Miami, Florida, Dec. 17-19, 1977) vol. I, 129-164, H. and P. Hollien (eds.), Amsterdam.

Fujimura, O. (1979b): "An analysis of English syllables as cores and affixes", Zs.f.Ph., Sign and system of language, Heft 4/5, 452-457.

Fujimura, O., T. Baer, and S. Niimi (1979): "A stereo-fiberscope with a magnetic interlens bridge for laryngeal observation", JASA 65, 478-480.

Fujimura, O. and Y. Kakita (1979): "Remarks on quantitative description of the lingual articulation", in Frontiers of speech communication research, S. Öhman and B. Lindblom (eds.), 17-24, London: Academic Press.

Fujimura, O. (forthcoming-a): "Elementary gestures and temporal organization -- What does an articulatory constraint mean?", Proc. of the International Symposium on the Cognitive Representation of Speech in their series 'Advances in Psychology', G. Stelmach and P. Vroon (eds.).

Fujimura, O. (forthcoming-b): "Fiberoptic observation and measurement of vocal fold movement", Paper presented at the Conference on the Assessment of Vocal Pathology, NIH, Bethesda, Maryland, April 17-19.

Fujisaki, H. (1977): "Functional models of articulatory and phonatory dynamics", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 347-366, Tokyo: University of Tokyo Press.

Harshman, R., P. Ladefoged, L. Goldstein, and J. Declark (1974): "Factors underlying the articulatory and acoustic structure of vowels", JASA 55, 385.

Hirano, M. (1977): "Structure and vibratory behavior of the vocal folds", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 13-30, Tokyo: University of Tokyo Press.

Hirose, H. and T. Gay (1972): "The activity of the intrinsic laryngeal muscles in voicing control -- an electromyographic study", Phonetica 25, 140-164.

Hirose, H., H. Yoshioka, and S. Niimi (1978): "A cross language study of laryngeal adjustment in consonant production", University of Tokyo AB RILP 12, 61-72.

Kent, R.D. and D. Minifie (1977): "Coarticulation in recent speech production models", JPh 5, 115-133.

Kiritani, S., K. Itoh, and O. Fujimura (1975): "Tongue pellet tracking by a computer-controlled X-ray microbeam system", JASA 57, 1516-1520.

Kiritani, S. and H. Imagawa (1976): "Principal component analysis of tongue pellet movement", University of Tokyo AB RILP 10, 15-18.

Kiritani, S., F. Miyawaki, O. Fujimura, and J.E. Miller (1976): "A computational model of the tongue", University of Tokyo AB RILP 10, 243-251.

Kiritani, S., S. Sekimoto, and H. Imagawa (1977): "Parameter description of the tongue movements for vowels", University of Tokyo AB RILP 11, 31-38.

Ladefoged, P.N. (1977): "The description of tongue shapes", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 209-222, Tokyo: University of Tokyo Press.

Liberman, M.Y. (1978): "Modeling of duration patterns in reiterant speech", in Linguistic variation, models and methods, D. Sankoff (ed.), 127-138, New York: Academic Press.

Liberman, M.Y. and L.A. Streeter (1978): "Use of nonsense-syllable mimicry in the study of prosodic phenomena", JASA 63, 231-233.

Liberman, M.Y. and A. Prince (1977): "On stress and linguistic rhythm", Linguistic Inquiry 8, 249-336.

Lindblom, B. (1963): "Spectrographic study of vowel reduction", JASA 35, 1773-1781.

Lindblom, B., R. McAllister, and J. Lubker (1977): "Compensatory articulation and the modeling of normal speech production behavior", in Articulatory modeling and phonetics, R. Carré, R. Descout, and M. Wajskop (eds.), 148-161.

Lovins, J.B., M.J. Macchi, and O. Fujimura (1979): "A demisyllable inventory for speech synthesis", in Speech communication papers, presented at the 97th Meeting of the Acoustical Society of America, J.J. Wolf and D.H. Klatt (eds.), 519-522.

MacNeilage, P. (1970): "The motor control of serial ordering of speech", Psychol. Rev. 77, 182-196.

Mattingly, I.G. (1977): "Syllable-based synthesis by rule", 9th International Congress on Acoustics, Madrid, July 4-9, 1977, Contributed papers 1, 512.

McCutcheon, M.J., S.G. Fletcher, and A. Hasegawa (1977): "Video-scanning system for measurement of lip and jaw motion", JASA 61, 1051-1055.

Mermelstein, P. (1973): "Articulatory model for the study of speech production", JASA 53, 1070-1082.

Miller, J.E. and O. Fujimura (1979): "A graphic display for combined presentation of acoustic and articulatory information", in Speech communication papers, presented at the 97th Meeting of the Acoustical Society of America, J.J. Wolf and D.H. Klatt (eds.), 221-224.

Minifie, F.D., C.A. Kelsey, J.A. Zagzebski, and T.W. King (1971): "Ultrasonic scans of the dorsal surface of the tongue", JASA 49, 1857-1860.

Muller, E.M. and J.H. Abbs (1979): "Strain gauge transduction of lip and jaw motion in the midsagittal plane: refinement of a prototype system", JASA 65, 481-486.

Nakajima, T. (1977): "Identification of dynamic articulatory model by acoustic analysis", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 251-275, Tokyo: University of Tokyo Press.

Nelson, W.L. (1979): "Automatic alignment of phonetic transcriptions of continuous speech utterances with corresponding speech-articulation data", in Speech communication papers, presented at the 97th Meeting of the Acoustical Society of America, J.J. Wolf and D.H. Klatt (eds.), 63-66.

Perkell, J.S. (1977): "Articulatory modeling, phonetic features and speech production strategies", in Articulatory modeling and phonetics, R. Carré, R. Descout, and M. Wajskop (eds.).

Pierrehumbert, J. (1979): "Intonation synthesis based on metrical grids", in Speech communication papers, presented at the 97th Meeting of the Acoustical Society of America, J.J. Wolf and D.H. Klatt (eds.), 523-526.

Sawashima, M. (1979): "A supplementary report on speech production", Proc.Phon. 9, vol. I, 49-56.

Sawashima, M. and H. Hirose (1968): "New laryngoscopic technique by use of fiber optics", JASA 43, 168-169.

Shirai, K. and M. Honda (1977): "Estimation of articulatory motion", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 279-302, Tokyo: University of Tokyo Press.

Sondhi, M.M. and B. Gopinath (1972): "Determination of vocal-tract shape from impulse response at the lips", JASA 49, 1867-1873.

Sonoda, Y. (1977): "A high sensitivity magnetometer for measuring the tongue point movements", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 145-156, Tokyo: University of Tokyo Press.

Sternberg, S., S. Monsell, R.L. Knoll, and C.E. Wright (1978): Information processing in motor control and learning, G.E. Stelmach (ed.), 117-152, Academic Press.

Stevens, K.N. (1972): "The quantal nature of speech: evidence from articulatory-acoustic data", in Human communication, A unified view, P.B. Denes and E.E. David (eds.), 51-66, New York: McGraw-Hill.

Titze, I.R. and D.T. Talkin (1979): "A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation", JASA 66, 60-74.

DISCUSSION

H. Hirose, M. Hirano, and J.S. Perkell opened the discussion.

H. Hirose emphasized that we have to be careful in the interpretation of the electromyographic data, in particular because the relationship between the degree of muscle contraction and EMG output - in the case of speech muscles - is linear only under very special conditions. In order to get some idea of the relationship between the EMG pattern and the articulatory events we have to combine several methods. As an example, Hirose showed some EMG and X-ray microbeam data recorded simultaneously. He concluded that modern methods for the investigation of speech production can also be applied to the analysis of pathological patterns of movements and furthermore perhaps help towards a better understanding of the role of related parts of the central nervous system in speech production.

M. Hirano discussed various techniques employed for the study of the morphology and function of the vocal folds. He demonstrated that there are two different kinds of fibrous components in the vocal folds, namely the elastic and the collagenous fibres and showed how the vocal folds consist of more layers, partly from a histological and partly from a mechanical point of view (cf. vol. I, p. 189).

J.S. Perkell said that from his point of view the use of movement and EMG data along with sophisticated physiological modelling is only in its infancy with respect to the contribution that these techniques will eventually make to our understanding of speech production, dynamics, and hopefully also control strategies.

Then he commented upon the need for additional and better data in the third dimension, i.e. the cross-sectional area function, to supplement good midsagittal data. The need for such data is illustrated by the range of current notions that we have about factors, which underlie or constrain vowel categories. For example, B. Lindblom and his co-workers have proposed that a vowel category may be determined by an interaction between perceptual distance and some measure of ease of articulation; M. Lindau has proposed a primary role for the acoustic factors; K. Stevens has suggested some role for the patterns of tongue-to-maxilla contact; S. Wood and others have suggested that the vowel categories are determined by quasi-discontinuous relationships between the place of constriction and the sensitivity of formants to changes in place and degree of this constriction, along with factors related to the muscular anatomy; and Fujimura, working with the tongue model, has suggested a role for discontinuous relationships between muscle contractions and area function. Perkell noted that we have no way of disproving any of these hypotheses, and it may well be the case that to some extent all of them are valid. He concluded that to begin to untangle all the possible influences on vowel categories, we need a lot more well controlled work to test each one of these hypotheses, and that improved knowledge of area functions along with other factors is obviously essential for the evaluation of all the hypotheses on articulatory correlates of sound categories.

Then Perkell turned to the question about the X-ray dosage for different X-ray techniques. Perkell and his co-workers have made some dosage measurements and he gave the following values for the dosage that the subject would get:
10 rads/min for 35 mm conventional cineradiographic film (60 frames/sec); 2,5 rads/min for 35 mm high speed film and for 60 mm conventional cineradiographic film; 600 mrads/min for 16 mm high speed cineradiographic film; 260 mrads/min for video-tape.

Perkell noted that the microbeam system rarely gets above approximately half of these values, but under most circumstances the microbeam exposes the subject to a much smaller dosage. Finally, Perkell mentioned that the X-ray unit they are using allows for simultaneous views in the anterior-posterior and in the lateral dimensions, and he hoped that they might be able to obtain information which will contribute to our insufficient knowledge of area functions.

O. Fujimura confirmed that the dosage for the X-ray microbeam system is about one half of the smallest dosage obtained with other X-ray systems, namely 120 mrads/min. But the frame rate used for the estimate of 120 mrad is 120 frames per sec., i.e. twice the rate that Perkell used for his estimate. And in order to derive the total energy absorbed into the body, the dosage given should be multiplied by the area under exposure; since the 120 mrads estimated for the microbeam system assumes a constant exposure over a small area of 1 cm^2, the product is obviously 120, whereas the exposed area is much larger in the case of the two other X-ray systems.

E. Keller found the ultra sound technique a valuable alternative to the X-ray technique. The great advantage is that the exposure time can be considerably longer than the exposure time using cineradiography. Keller also pointed out that the frame rate is limited with the X-ray methods, which is a problem if we want to make measurements of speed of articulation, for instance. Finally, Keller said that with the ultrasonic method using a scanning beam, i.e. a system where a beam is sent back and forth several thousand times per sec., the whole surface of the tongue can be recorded, for instance, contrary to what can be obtained by a single beam system.

O. Fujimura claimed that for the X-ray microbeam system the net total of exposure time for one session is typically about 10 min., and often they run two or three sessions per subject. The total dosage given to the subject in terms of energy absorbed in one session is comparable to the amount of dosage one gets from the cosmic rays during one year. Concerning the limitation of frame rate, Fujimura replied that if one is interested in studying very fast movements in one portion of the tongue, which is the normal application of the ultrasonic technique, the number of pellets can be reduced and thereby a frame rate of up to 1000 frames per sec. can be obtained, so the frame rate for the microbeam system is not restricted to anything like 120 frames/sec.

Finally, Fujimura mentioned that for the velum height measurements - using the X-ray microbeam system - the pellet is not glued directly on to the velum as is the case with tongue pellets. Instead, a narrow strip of a very flexible plastic sheet is inserted through the nostril, covering the pellet, and this keeps the pellet in position, in contact with the upper surface of the velum.

J. Ohala emphasized that the estimates of radiographic dosage that we find in the literature vary tremendously. Furthermore, he referred to a study revealing an increased incidence of cancer in the thyroid from a population who had been radiated 30 years ago as children, with a dose of 6 rads, but these cancers did not develop until now. Ohala concluded that though the vocal tract is very important for us, we have to be very cautious in estimating our dosages. He advocated an intensification of our search for alternative ways of getting vocal tract informations.

G. Fant mentioned the possibility of measuring the impedance between two points, e.g. the upper and lower lip, as an alternative method for tracking the dynamics of articulation.

H. Künzel mentioned a very simple instrument for real-time recording of velar elevation, developed at the Institute of Phonetics in Kiel. The system consists of an optical probe - with an outer diameter of 3 mm - inserted through the nostril. The probe emits light which is reflected as a function of velar elevation. The linear function of the system has been controlled by simultaneous X-ray recordings.

C. Scully mentioned another approach, which works back from the aerodynamic stage and infers movements of the articulators from aerodynamic data. Such a technique can give us some idea of the size of the constrictor across which a pressure drop can be measured. What sort of range and what degree of accuracy this yields is an open question at the moment, but it is being investigated.

O. Fujimura mentioned a new technique, suggested by Dr. Sinada, where the pellet position is detected purely magnetically. The only disadvantage is that at the moment only one pellet can be tracked.

The indirect methods are very useful, in particular for practical purposes like clinical applications, training of articulatory gestures, and so on. But they need calibration and here the microbeam system could also be used.

S. Smith claimed that the electroglottographic method tells us something about the state of the musculature, i.e. whether it is relaxed or contracted.

O. Fujimura said that a technique for measuring the state of the muscular contraction by some physical means would be advantageous if we can establish a way to calibrate it.

CORTICAL ACTIVITY IN LEFT AND RIGHT HEMISPHERE DURING LANGUAGE RELATED BRAIN FUNCTIONS

Niels A. Lassen and Bo Larsen,[1] Department of clinical physiology, Bispebjerg Hospital, DK-2400 Copenhagen, Denmark

Chairpersons: Peter Ladefoged and Hans Günther Tillmann

The blood flow through the brain cortex varies with the functional state of the tissue. Just as in skeletal muscle or in various glands, an enhanced level of nerve cell activity invokes an increase in tissue metabolism and in blood flow. Thus, it was found by Olesen from our group that rhythmical movement of the hand augments regional blood flow in the contralateral central (hand) cortex by 20 to 30 per cent. It was subsequently verified that indeed not only flow but also oxygen uptake is increased in that same area during hand exercise. We have used regional blood flow measurements to map the cortical areas active in various types of language related brain functions. A summary of our findings will be given.

The method used for measurement of regional cerebral blood flow in man

The radioactive isotope Xenon-133 is used. It is produced in a nuclear reactor as a split product of uranium. Like the non-radioactive Xenon isotopes, Xe-133 is an inert gas and (like nitrogen, N_2) it does not react chemically with any molecules in the body. It is simply distributed according to the tissues' solubility. We use it in the form of a physical solution in saline in a dose of approximately 5 MilliCuries per injection (1.5 ml). The radiation exposure is negligible; it is much less than that of a single conventional X-ray study. This means that a series of repeated injections with an interval of 15 minutes can be made in the same setting without any radiation hazard. We take advantage of this by usually performing a series of 4 or 5 injections in one study: first at rest and then during a series of different forms of brain work - in this case involving various language related types of brain functions.

The Xenon-133 containing sterile saline is injected into a big artery on one side of the neck, the internal carotid artery. It supplies the anterior 3/4 of the brain (usually the posterior

---

1) The paper was given by N.A. Lassen.

part of the brain, the occipital lobe's inner side, is not receiving the isotope by this injection as its arterial supply comes from a different artery, namely from the vertebral artery). With each internal carotid supplying (normally) only the ipsilateral cerebral hemisphere, and by injecting only one side, we obtain maps of blood flow distribution in one hemisphere only. This is a distinct limitation with regard to studying hemispheric differences: we have to rely on comparing a series of left hemisphere observations with those on the right side in other subjects, and cannot in the same subject observe both sides simultaneously.

Using a special isotope camera with 254 small detectors, we observe the arrival and subsequent wash-out of the Xenon-133 in regions of the size of approximately 1 cm^2. The tissue element "seen" has the form of a cone traversing the injected hemisphere. Due to absorption of radiation it is, however, the superficial cortex we see best. The regional blood flow is calculated from the slope of the Xenon-133 wash-out curve during the first minute following the injection of the radioactive bolus (that takes only one second). When a test is performed, such as counting or reading, the subject is asked to start performing approximately 10 seconds before the Xenon-133 injection, and then continue for 60 seconds (the injection is not felt by the subject). The interval of approximately 15 minutes between injections is necessary in order to clear the brain of radioactivity before injecting the next dose (we can actually use a shorter interval, and then compensate for remaining radioactivity).

The technique is not entirely atraumatic: it involves the cannulation and injection into the blood flowing to the brain and a risk of compromizing this flow exists. We have not encountered any complications in the series of 350 subjects studied in our laboratory (over a period of 4 years) with the technique described here. Yet, this risk restricts us to study patients with neurological symptoms in whom cerebral angiography is indicated, i.e. in whom a cannula is placed in the carotid artery for X-ray study. This means that normal subjects cannot be studied. Nevertheless, our series of neurological patients comprises cases without focal tissue abnormalities (patients studied because of arterial aneurisms or because of an epileptic seizure, cases of suspected brain tumor, etc.). The results obtained in such cases (approximately 20% of our patients) constitute our equivalent of normal man. The main part of the studies reported below pertain to such "normal" cases. The consistency of the results leaves no doubt that the data may indeed be taken to pertain to normal man.

Results

A. The awake resting state. With closed eyes in a darkened silent room and completely at rest the normal pattern of blood flow distribution shows the highest values in the frontal lobe (approximately 10% above the hemispheric mean).

B. Listening to words. Simple noise produced with Barany noise apparatus increases flow in the hearing cortex only minimally. Listening to sounds (Seashore test) or onomatopoetica as "crack", "bang", "whiz", on the other hand, clearly activates this area on both sides (15-30% increase in flow). The area comprises Wernicke's center of language on the left side (all our subjects were right handed). Listening to music caused the same effect. Our data do not suggest a hemispheric difference with these two forms of simple listening tests.

Listening to more complex spoken language produces increased flow on the left side. But since this area overlies the basal ganglia and since a flow increase here is often seen with the unspecific more global flow increase accompanying increased attention, we cannot assert the specificity of this activation.

C. Talking. Automatic talk in the form of counting repeatedly to twenty at a rate of one digit per second activates the hearing cortex, the primary (rolandic) mouth area and the supplementary motor area.

All these changes are bilateral. The pattern tends to be less sharply demarcated on the right side than on the left.

Word naming in the form of finding words of 5 flowers, 5 types of furniture, etc., activated the same three areas and caused a constant activation of the whole prefrontal region as well (cf. the comments made under reading aloud and internal speech).

D. Reading aloud. This activates six areas in both hemispheres. In addition to the three areas seen during automatic talk, the following areas also become active: the visual association cortex in the posterior part of the brain, the frontal eye field that often merges with the mouth area, and the low-posterior part of the frontal lobe (with Broca's area on the left side) which we commented on above.

We cannot - in most of our cases - see the primary visual cortex as it is usually supplied by the vertebral artery. But from animal studies it is evident that this area becomes more active during visual stimulation. Hence, including this area, a total of fourteen discrete cortical areas, seven on each side, are active during reading aloud. Often, the prefrontal cortex anterior to the supplementary motor area, is also activated.

Reading a text aloud is a prime example of the fundamental mode of operation of the cerebral cortex in performing complex tasks (and there are probably no simple ones!): collaboration between discrete cortical areas, each performing a specific job. It is the pattern of activation that is related to the task, not any single area. There are, in other words, no isolated center solely responsible for solving a complex task as also emphasized by the late Alexander R. Luria.

So far we have not been able to discern individual patterns of cortical activity of such a nature as to suggest fundamental differences between individuals.

The role of the right hemisphere in this complex language function is not clear. But data from the literature suggest that production and analysis of language melody and perhaps even of gestures related to speaking may predominantly reside on the right side.

E. Reading silently. If the same subject after reading aloud reads silently, the change in the map of blood flow, compared to that at rest, is particularly easy to interpret: then the primary sensori-motor mouth area and the auditory cortex do not become active. All the other areas are, however, seen to be active.

F. Internal speech. Memorizing the text internally causes a small increase in the mean blood flow, predominantly in the frontal lobe (often especially in the prefrontal cortex).

This type of "global" activation is seen with any task that the subject makes an intellectual effort in accomplishing.

In our opinion, internal speech is a mental function which is just as real as love, hate, or memories. It is a solid fact of introspection. This is supported by the fact that one can readily think in different languages. But, while asserting the psychological reality of internal language, we would consider it imprudent to follow Luria's speculation that this language function has a special grammatical construction.

It is tempting to revert the argument and to state that internal speech is the only true or "essential" language function.

How about a patient paralyzed by a disease or Curare and who tries to speak? Can one state that he has no language function? In a way, all the external manifestations of language functions are non-essential - solely the internal functions of language understanding and production - both comprized in the concept of internal speech - are truly essential.

G. Aphasia. We have studied the blood flow map on the left side in a series of classical aphasia patients, mostly cerebro-vascular accidents ("apoplexy"). Confirming well-known facts, the "fluent" aphasia cases had defects (on the flow map) in the posterior speech area of Wernicke, "non-fluent" or "motor" aphasia had defects in the primary mouth area (sometimes but not always extending to Broca's area), "global" aphasia had large defects covering both Wernicke's area and the mouth area. No studies were made on the right hemisphere.

H. Auditory agnosia (comprising word deafness). A rare case of sensory aphasia due to bilateral temporal lobe infarcts was studied in some detail. The patient, a 63 year old man, first suffered an attack of mild fluent aphasia, lasting one month. Some months later, he suddenly lost all ability to understand any spoken words and had some difficulty in recognizing non-verbal sounds. Yet, his threshold for perceiving pure tones was normal for his age. In other words, he was not deaf. But he could no identify any words. Not even his own name, or simple words, such as yes and no. All other language functions were intact: talking, reading, writing. This state is in neurological terminology called "auditory agnosia".

Specialized investigations suggested that an acute right-sided lesion of the hearing cortex had cut off ("disconnected") the remaining posterior part of the left superior temporal gyrus (Wernicke's center) from its remaining input (that from the left side having been destroyed by the first stroke). Computerized tomography (CT-scanning) showed the bi-temporal infarcts as hypodense lesions involving Heschl's gyrus bilaterally. Regional blood flow studies during listening to sounds showed no activation of the upper part of the left temporal lobe area (Wernicke's area): The sound analyzer was not turned on!

This case is interesting for three reasons:

1) A right hemisphere lesion (lesion no. 2) produced a massive language handling defect in a right-handed subject. There are other cases of this type recorded in the literature.

2) Preservation of normal speech in a subject deprived of all meaningful auditory feed-back. The fact that completely deaf patients can speak fairly normally also confirms that some of the speculations concerning the necessity of the normal auditory feed-back for speech production have been exaggerated. The importance of this feed-back for language acquisition is not questioned in this argumentation.

3) The patient had completely normal early components in his auditory evoked response. Hence this response cannot originate in the primary auditory cortex -Heschl's gyri- as these were massively destroyed bilaterally.

Concluding comments

Many linguistic and phonetic problems related to cortical function could be posed in relation to the findings we have summarized in this paper. However, it is appropriate here to stress the poor temporal resolution (1 minute) and spatial resolution (1 cm^2 with superposition of deeper layers) involved in the registration of the regional blood flow. Certainly, we cannot by this approach say anything about the detailed way in which the cortical areas collaborate in language functions.

It surprised us to find that a simple sound-rhythm discrimination test (Seashore) activated the auditory association cortex to much the same extent as do music or language. Apparently, the whole sound analyzer works as a unit.

The major finding was in our opinion the bilateral and practically symmetrical cortical involvement in all language functions. The possibility of a special role of the right side for prosody is mentioned. We have no data pertaining specifically to this point.

A comment on memory may be appropriate. It appears that this function is disseminated in the brain: visual memory in the visual association cortex, tactile memory in the sensory cortex, etc. Thus it is not surprising that word memory resides in the auditory association cortex in the temporal lobe. That it is predominantly on the left side is, however, completely mysterious! Could it be that the speed of language perception (and production) precludes major inter-hemispheric information exchange in this most human or "highest" of all types of brain work?

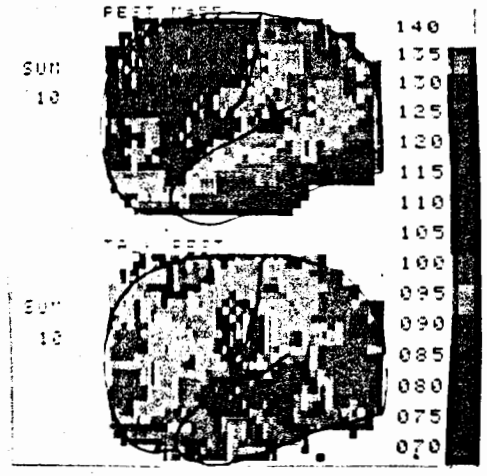


Fig. 1 Intact normal man, regional cerebral blood flow, rCBF, map, left hemisphere.
The original illustration is in colours and therefore the black and white reproduction has distorted the scale. The legend to this figure gives a verbal description of the increase in flow clearly seen on the original and also visible on this reproduction.
The upper frame shows the rCBF map at rest, average picture of 10 cases. The map is expressed in percent flow deviation from the mean hemispheric value (averaging 55 ml/100g/min in these cases).
The lower frame shows the average rCBF map during automatic speech expressed as percentage deviation from the map at rest. Three areas of consistent flow increase ("activation") are seen (in this black and white reproduction the areas are slightly darker than the rest, with still darker edges): the supplementary motor area (at the top), the sensory-motor mouth area (upper mid), and the posterior part of the superior temporal gyrus (lower mid, Heschl's gyri and Wernicke's area).
Changes in the right rCBF map during automatic speech are practically the same. Broca's area is usually seen with fluent speech.

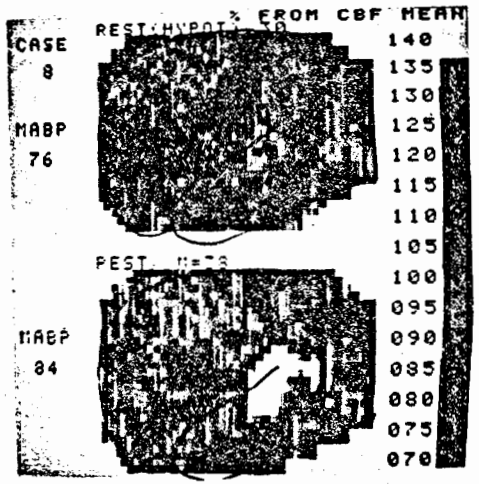


Fig. 2 Stroke case with aphasia (case 8 of our series), regional cerebral blood flow map, left hemisphere.
The upper frame shows rCBF map at rest during normotension (mean arterial blood pressure 84, mean rCBF 38 ml/100g/min).
Note the dramatic increase of flow in Wernicke's area (white) as flow rises: Luxury perfusion 8 days after onset of stroke, probably overlying an infarct, with abnormal pressure passive flow regulation.

References

Larsen, B., E. Skinhøj, and N.A. Lassen (1978): "Variations in regional cortical blood flow in the right and left hemispheres during automatic speech", Brain 101, part 2, 193-209.

Lassen, N.A., D.H. Ingvar, and E. Skinhøj (1978): "Brain function and blood flow. Changes in the amount of blood flowing in areas of the human cerebral cortex, reflecting changes in the activity of those areas, are graphically revealed with the aid of radioactive isotopes", Scientific American 239, 62-71.

Soh, K., B. Larsen, E. Skinhøj, and N.A. Lassen (1978): "Regional cerebral blood flow in aphasia", Arch. Neurology 35, 625-632.

Lassen, N.A. (1978): "Cerebral blood flow in cerebral ischemia. A review", European Neurology 17, suppl. 1, 4-8.

DISCUSSION

Victoria Fromkin, Michael Studdert-Kennedy and Peter MacNeilage opened the discussion.

Victoria Fromkin quoted Fournier's statement (in the late 19th century): "Speech is the only window through which the physiologist can view the cerebral life" and added that it should also be recognized that the brain is a window through which we will be able to observe the linguistic life.

Victoria Fromkin then expressed her hope that these new techniques would reveal to what degree language is a special function of the brain rather than a particular case of more general faculties. We ought to find out whether patients show differences in brain activity when they are subjected to stimuli of varying degrees of phonetic or linguistic complexity. And she mentioned that there might be different reactions to known versus unknown language stimuli, which again might be different from clearly non-linguistic input. Finally, different reactions might also be expected from patients automatically repeating memorized formulae rather than producing or reacting to free, creative speech.

In connection with the supposedly unexpected activity of the right hemisphere during automatic speech Victoria Fromkin mentioned that it is well known that even people who display a marked hemispheric specialization will always show some activity even in their right hemisphere during speech.

Victoria Fromkin further pointed to the dangers of drawing too far-reaching conclusions from observations based solely on patients with abnormalities of the brain. And she stressed the importance of looking for a convergence of results from different techniques.

Finally, she mentioned the importance of sensory aphasia cases, such as had been described in the lecture, for the debate on whether grammar exists apart from perception and production.

Michael Studdert-Kennedy: I think it is quite clear that techniques of this kind, such as the blood flow techniques, the more advanced analyses through EEG work, and perhaps the development of cooling techniques for isolating parts of the brain in the normal brain, are going to be much more important in the future than the type of behavioural studies that we have had to rely on in the past.

Michael Studdert-Kennedy then mentioned the problem raised by Victoria Fromkin in that these techniques are always used with patients. And he pointed out that in cases of apoplexy the right hemisphere could slowly be taking over functions normally performed by the left hemisphere. Therefore these new techniques should be developed and be made usable with normals.

He then continued: Obviously, the finding that there is a large amount of right hemisphere activity as well as left hemisphere activity is not a surprise. Because presumably there is a coordination of function between the two sides of the brain.

Nonetheless, there are certain properties of one side of the brain rather than the other that do arouse interest. And that seems to me to be important in understanding the nature of linguistic communication. I am referring here particularly to the famous relationship between speech and handedness.

It seems to me that an understanding of that relationship would take us rather a long way to understanding what the prior signalling conditions are for communication.

In this regard I think that the current developments in the work on sign language is tremendously important. Because it does seem that a prerequisite for linguistic communication is a motor system that is capable of very fine, rapid articulation.

One has only got to ask oneself what sort of a sign language could be developed if one was forced to use one's feet to realize that one absolutely has to have pieces of machinery that can be moved very, very fast. And so the motor control of that machinery which appears to be in some way common between the speech mechanism and the hand mechanisms, are of great interest. And I think that one very exciting possibility that these techniques look forward to is an elaboration and an understanding of these links.

In that respect I wonder, too, what the prospects are for looking at these processes developmentally.

Michael Studdert-Kennedy then drew attention to the sensory-motor integration functions described by Niels Lassen, particularly those concerned with speech and hand movements. These integration functions would appear to be a necessary prerequisite for the development of language. And he mentioned how children exposed to sign language will start imitating this at the same time as spoken language is normally developed.

Finally, he, too, suggested that the new techniques be used with different types of linguistic stimuli.

Peter MacNeilage was particularly interested in the spreading of activity from the temporal to the parietal lobes, having observed in the slides an upward spreading of activity in the parietal lobe during reading as compared to counting, and a still further spreading during listening. And he continued: The reason I am interested in the parietal lobe is its involvement in what is usually called conduction aphasia, and because I believe at the moment that the posterior and inferior parietal lobe is of some importance in the formulation of complex, voluntary movements.

Peter MacNeilage, too, mentioned the possibility of reorganization of brain functions after hemispheric lesions. And he suggested in the specific case mentioned by Lassen that possible simultaneous damage to Heschl's gyrus should be considered.

He then said: You may not exactly have intended to say this. But when you were talking about the finger movement task, you pointed to the fact that there was a rather circumscribed and small area of high activity in area four, that did not extend very anteriorly. On the other hand, there was a much larger and more widespread area of activity in the somatic-sensory cortex. And I believe you said that you thought the somatic-sensory activity was of more importance than the motor activity. I would like you to clarify this remark. [Niels Lassen: "That is correct."] Because it seems to me to relate to a rather general question about the extent to which we can simply assume a linear relation between the amount of activity, or wideness of distribution of activity, and the importance of the function.

It seems from my point of view that there may be parts of the cortex that can get their job done with less blood flow than other parts.

In your Scientific American paper you talked a little more of the role of the supplementary motor cortex than you have here. You still believe that the supplementary motor cortex has an important organising role in the production of speech? Because an alternative hypothesis is possible, namely that it simply has to do with initiation, or facilitation, of action in a rather general sense.

One could possibly argue that it has the equivalent of an attentional role on the motor side. It facilitates things happening without actually having much to do with the details of the control function themselves.

Coming back to the question of skilled, voluntary movement, I would like to ask to what extent you have studied unskilled voluntary control versus skilled voluntary control. That is in particular in relation to learning a skilled voluntary task.

And finally, I have heard that Brenda Milner, using sodium amytal studies, has shown a rather interesting relationship between the controlling hemisphere in left handers for speech, and for skilled voluntary movement of other kinds.

Niels Lassen, answering the first three discussants, said how surprised he and his colleagues had been when they saw to what degree the entire auditory association cortex was activated when a patient was stimulated with even very simple sounds. Stimulation with longer sequences and more complex stimuli was found to raise the level of activity a little more, but in the same area. Since the difference in reaction to simple and complex stimuli was found to be so small, the general rise in activity may be thought of as a sort of local attention phenomenon, where the whole auditory system is activated by any incoming signal.

Concerning the proposals to use more differentiated stimuli, Niels Lassen mentioned that he had received a stimulus tape from the Phonetics Institute, Copenhagen, containing white noise, isolated vowels, simple CV-syllables as well as connected speech, and was planning to use this in further experiments.

About the questions concerning the supplementary motor area, Niels Lassen said that he had found this area particularly active during complex movements. The relatively high level of activity in this area during speech could therefore be explained by the fact that speech is produced by very fast and complex movements.

As to the question about Heschl's gyrus, Niels Lassen said that there were damages to that area on both sides, but that he was not sure whether it had been completely destroyed. What was clear was that there no longer arrived any information to the auditory association cortex on the left side. That was evident from the lack of flow increase in that area when listening to words.

Niels Lassen confirmed that the parietal lobe is indeed very active, also during speech. But this appears to be part of the general arousal of the brain, since this area becomes active with any kind of activity on the part of the patient.

Barbara Prohovnik mentioned that people in Lund were using Xenon inhalation methods, which are non-invasive, to obtain similar traces.

Niels Lassen answered that the inhalation methods gave less well defined results because of limitations in the time constants of these methods.

Prompted by several people, Niels Lassen stated that the method he described measures the average activity of the brain over at least ten or fifteen seconds. Generally, a recording averages over the first thirty or forty seconds where the important information is concentrated. It is known from animal studies that there is a time lag of two or three seconds from the time of the injection to the time when changes in the blood flow can be clearly detected, and it disappears over ten to fifteen seconds. A new injection may be made after about three minutes. But successive recordings have even been made with only one minute intervals.

Vincent van Heuven suggested that we look not only for the areas of increased blood flow, but that we also examine what areas are inactive during a particular task. Both Vincent van Heuven and Niels Lassen commented on the fact that the brain always shows activity somewhere, even when the patient is at rest. But Niels Lassen said that he had observed not only increases but even reductions in the level of activity in certain areas when the level rose in other areas because the patient concentrated on a particular task.

John Laver was sceptical about the reported case of a bilateral lesion which had made auditory feed-back impossible. Experience shows that this should cause a progressive deterioration of the articulatory accuracy of the patient's speech, which apparently it had failed to do in this case.

H. Mol mentioned that he knew of a totally blind and deaf man, who has excellent speech performance. His deafness developed suddenly at the age of 31 as the result of meningitis.

Niels Lassen said that this case, just as his own, strongly supports the notion of the unimportance of auditory feed-back for speaking a well established language.

NEW METHODS OF ANALYSIS IN SPEECH ACOUSTICS

Hisashi Wakita, Speech Communications Research Laboratory Inc., 806 West Adams Boulevard, Los Angeles, California 90007, U.S.A.

Chairperson: Hans Werner Strube

Introduction

The recent development in digital techniques has brought substantial innovations to methods and techniques for acoustical analysis of speech sounds. The advantages of using digital computers over the conventional analog techniques are that the analysis processes can be repeated precisely and that the control of the parameters is relatively easy. The use of the digital computer also permits the processing of a large amount of data within a relatively short period of time with satisfactory accuracy. Because of the above advantages, digital techniques are playing a more and more important role in speech research. As this tendency becomes stronger, proper care has to be taken when the digital techniques are applied to speech research. This paper, thus, concerns primarily the recent digital techniques in the acoustic analysis of speech, particularly the linear prediction method, with special attention to its advantages and disadvantages, and also to the limitations involved in the technique.

The concept of linear prediction was first applied to speech analysis by Itakura and Saito in Japan (1966) and by Atal and Schroeder in the United States (1967). Since then the linear prediction method has been fairly thoroughly studied theoretically and experimentally (see Makhoul 1975; Markel and Gray 1976; Wakita 1976), and the method is currently being used as a powerful tool for acoustical analysis of speech sounds.

Linear prediction of speech

A very simplistic model of speech production as shown in Figure 1 (a) is assumed in the linear prediction of speech. The excitation source is an impulse and the filter, which mainly represents the vocal tract, has the frequency characteristics of resonances only, without any anti-resonances. The model thus exclusively represents the voiced and non-nasalized sounds.

For an analysis model, an inverse filter is assumed, which maintains the precise inverse relation between the input and the output of the production model, as shown in Figure 1 (b). Thus,

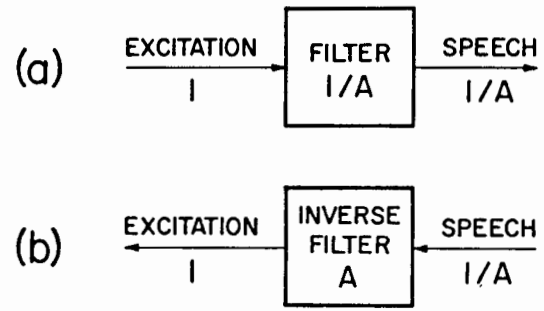


Figure 1. Models for the linear prediction method: (a) Production model; (b) Analysis model.

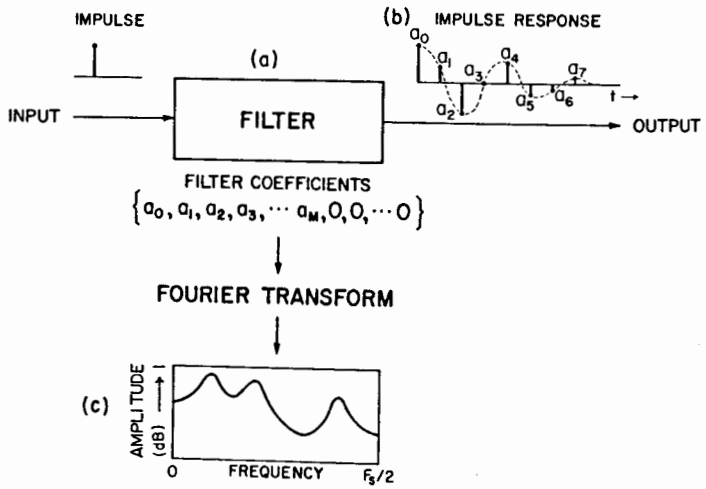


Figure 2. Determination of filter characteristics: (a) a model; (b) discrete impulse response; (c) frequency characteristics (transfer function) of the filter.

the problem in linear prediction analysis is to determine the characteristics of the inverse filter from a given input speech wave.

Since the linear prediction method is a digital technique, all the data, and parameters to specify the filter characteristics, are handled in a discrete sampled format instead of as continuous quantities. The main task of linear prediction is to predict the current speech sample $\hat{x}_n$ in terms of a linear combination of the past M samples. Letting the predicted current sample be $\hat{x}_n$, $\hat{x}_n$ is given by

$$\hat{x}_n = \alpha_1 x_{n-1} + \alpha_2 x_{n-2} + \ldots + \alpha_M x_{n-M} \qquad (1)$$

In equation (1), the α_i's are called predictor coefficients. They play a role of "weighting" the past samples to predict the current one. The problem in the linear prediction method is to determine these predictor coefficients in such a way so as to minimize the error between the current sample and the predicted one, and to relate the predictor coefficients to the parameters of the inverse filter. In this case, the sum of the squared errors over a certain period,

$$E = \sum_{n=1}^{N} (x_n - \hat{x}_n)^2 \qquad (2)$$

is minimized. Because of this, speech samples during this period are assumed to be sufficiently stationary so that the predictor coefficients do not change during this period.

How are the predictor coefficients thus determined related to physically meaningful parameters, that is, to the inverse filter in Figure 1 (b)? In general, the frequency characteristics of a filter can be determined by observing its impulse response when an impulse signal is applied to the filter as shown in Figure 2 (a). In the discrete case, the impulse response of a filter is then given as shown in Figure 2 (b). The amplitude at each sampled point in the impulse response is given by a_i and the period between the two sample points is given by the sampling period T. From this impulse response, the transfer function, A(z), of the filter is given by use of "z-transform" notation as

$$A(z) = a_0 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_M z^{-M} \qquad (3)$$

Equation (3) represents not only the transfer function of the filter but also the impulse response in the time domain. The a_i's in equation (3) are called filter coefficients. It is easily seen from Figure 2 (b) that the interpretation of the "z-transform" notation is that z^{-1} represents a unit delay in the time domain in terms of the sampling period T. Thus, the power of z^{-1} in equation (3) denotes the number of time delays.

Since $z=\exp(j2\pi fT)$, where j is the imaginary unit ($j=+\sqrt{-1}$) and f is frequency, equation (3) itself represents the discrete Fourier transform of the impulse response. Thus the frequency domain representation of equation (3) is given by applying the Fourier transform to the filter coefficients. In this case, the impulse response is truncated at t=MT and normally sufficient zeroes (e.g. 256 minus M zeroes) are added to the a_i's to ensure sufficient frequency resolution before the Fourier transform is applied. An example of a power spectrum obtained from the output of the Fourier transform is given in Figure 2 (c). Note that the frequency band is bounded at $F_s/2$ where $F_s=1/T$ is the sampling frequency. Note also that when the amplitude of the frequency components is represented on a logarithmic scale, the frequency characteristics of the inverse filter as shown in Figure 2 (c) become those of the vocal tract filter in Figure 1 (a) just by relabeling the negative sign of the ordinate with a positive sign.

One of the important features of the linear prediction method is that the predictor coefficients in linear prediction of speech can be shown to be identical to the filter coefficients with $a_0=1$. Consequently, minimizing the overall error in linear prediction is equivalent to finding the transfer function of the inverse filter of the analysis model in Figure 1 (b).

Analysis condition

Proper analysis conditions for the linear prediction method are important to ensure satisfactory results. The analysis conditions to be noted are (1) sampling frequency, (2) the number of coefficients, (3) time window and length, (4) window shift, and (5) preemphasis. The sampling frequency determines the frequency range of interest. The frequency range must be less than or equal to half the sampling frequency (normally the latter is chosen). The number of coefficients is dependent on the frequency range to be chosen. When the frequency range is exactly half the sampling frequency (F_s kHz), a good rule of thumb for the number of filter coefficients is from F_s+2 to F_s+4. The reason for this appears to be that there will be about $F_s/2$ resonances in the frequency band limited by $F_s/2$, provided that F_s is given in units of 1 kHz. Each resonance requires 2 coefficients for its representation, and so about F_s coefficients will be needed to account for the expected resonances in the analysis band. In addition, 2-4 coefficients are normally used for approximating the spectral slope due to the excitation source.

The analysis conditions (3) and (4) vary depending upon which of two different methods of linear prediction is used, the autocorrelation method or covariance method (e.g. Markel and Gray 1976). The two methods use different definitions for computing the coefficients from sampled speech. The autocorrelation method requires a window length of at least 1.5 pitch periods and a Hamming window is recommended to suppress the spectral disturbances in the high frequency region due to the edge effect of the time window. The covariance method, on the other hand, does not require any particular time window, and the window length can be less than a pitch period. Thus this method can be used for pitch-synchronous analysis of speech sounds. When a window length of less than a pitch period is chosen, care must be taken since the analysis results vary depending upon what portion of the pitch period is chosen for analysis. This method is particularly useful for extracting the true vocal tract characteristics by choosing the glottis-closed portion of the speech waves. The major disadvantage of the covariance method is that there is theoretically no guarantee for obtaining a stable transfer function for the inverse filter, and thus a more sophisticated algorithm is required to automatically process the cases of instability. Also a more sophisticated algorithm is needed for automatically windowing the speech wave into pitch-synchronous intervals.

The window shift in the covariance method, thus, involves a more complicated procedure than it does in the autocorrelation method. In the latter method, the window shift is rather arbitrary, depending upon the speech samples to be analyzed. The shift can be greater than the window length for steady-state sounds, whereas, for speech sounds in which the formant frequencies are rapidly changing, a smaller window shift will be better for obtaining the smooth contour of the formant frequencies.

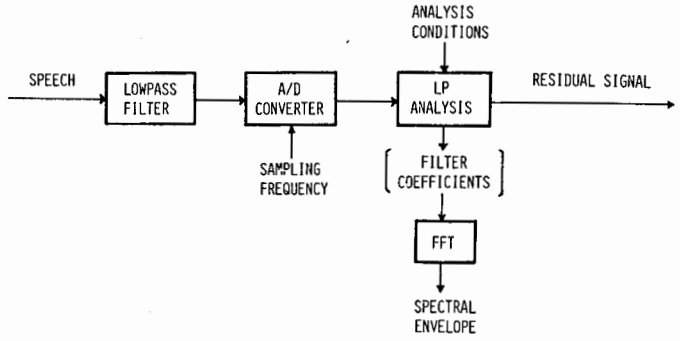


Figure 3. A block diagram to compute the smooth spectral envelopes of speech sounds by the linear prediction method.

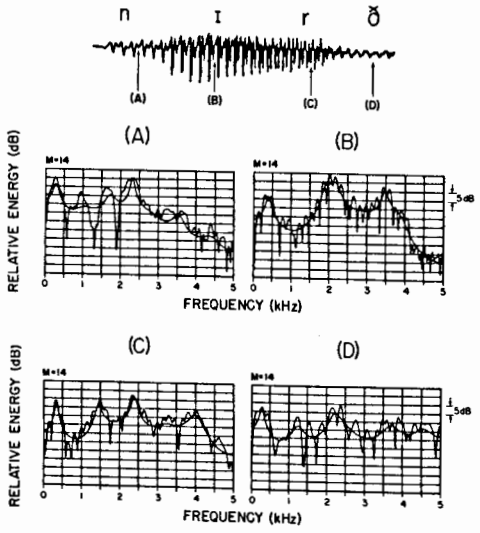


Figure 4. An example of linear prediction analysis. (Sampling frequency 10 kHz; number of coefficients 14; window size 20ms with a Hamming window and +6dB/octave preemphasis.)

A 6 dB/octave preemphasis is recommended for formant analysis. This is accomplished by taking the backward differencing of the sampled speech. The purpose of the preemphasis is to enhance the spectral peaks in the high frequency region. The 6 dB/octave preemphasis also roughly compensates the -12 dB/octave glottal source characteristics and the +6 dB/octave lip radiation characteristics.

Estimation of formant frequencies

As mentioned before, the Fourier transform of the predictor coefficients gives the frequency characteristics of the inverse filter, the inverse of which are the frequency characteristics of the vocal tract filter. Thus the procedure for obtaining the smooth spectral envelope by use of the linear prediction method is given by the block diagram shown in Figure 3. The speech signal is first digitized at some sampling frequency after being passed through a lowpass filter to limit the frequency band according to the sampling frequency. Linear prediction analysis is then performed using predetermined analysis conditions, and resulting in a set of filter coefficients for each speech segment analyzed. Smooth spectral envelopes are computed from the output of the Fourier transform of the filter coefficients with added zeroes. As a result of linear prediction analysis, the residual signal, which is an error signal given by equation (2) is saved for detecting pitch periods as will be described later.

An example of analysis results is shown in Figure 4. This example is a part of a sentence "Near the boat ..." and the spectral envelope estimation for /n/, /ɪ/, /r/, and /ð/ are shown in the figure together with the direct Fourier transform of the corresponding speech waves. It is seen that spectral peaks are well approximated by the extracted spectral envelope. However, the spectral dips due to anti-resonances as in the sound /n/ are ignored in the linear prediction method, in which the nasal tract is not considered. It should be noted that the linear prediction method was developed as a method for efficient speech analysis-synthesis telephony on the basis of the fact that the human ear is insensitive to spectral dips. Thus ignorance of spectral dips is not a major problem as far as analysis-synthesis telephony is concerned. However, if one is interested in more accurate estimation of spectral dips as well as peaks, a new model has to be developed, which is currently being investigated by some researchers.

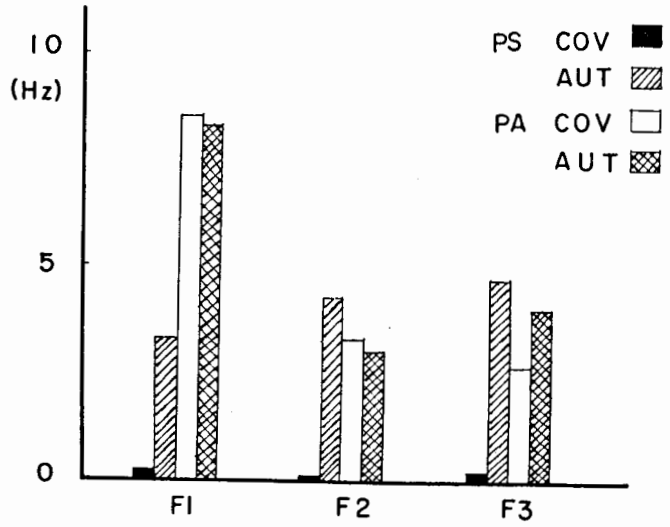


Figure 5. Evaluation of formant frequency estimation by autocorrelation and covariance methods for pitch-synchronous and pitch-asynchronous cases.

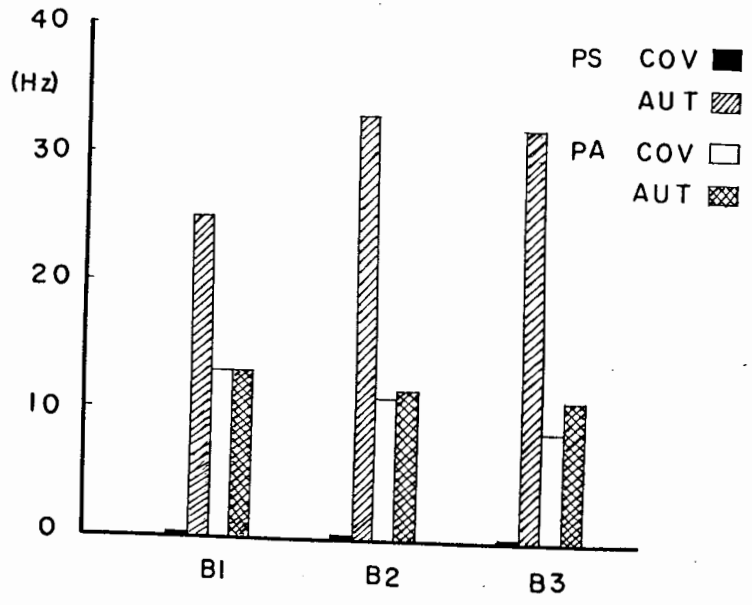


Figure 6. Evaluation of formant bandwidth estimation by autocorrelation and covariance methods for pitch-synchronous and pitch-asynchronous cases.

The formant frequencies are estimated from the smooth spectral envelope by finding the locations of the spectral peaks by a peak-picking method. Although this method is simple and worthwhile, it presents problems when two peaks are close together or merged into a broad peak. Another method is to compute the exact locations of the peaks by solving for the roots of the transfer function, A(z), of the inverse filter. In both methods, the spectral peaks do not always correspond to the formant frequencies, and thus a certain algorithm to automatically select formant peaks has to be designed (e.g. McCandless 1974). For both methods, a careful inspection of the analysis results is recommended before further processing of the formant frequencies is initiated.

Accuracy of formant estimation

It is rather difficult to determine the accuracy of formant estimation for natural utterances, since there is no way of accurately measuring the vocal tract configuration to compute its resonances while a sound is being produced. Chandra and Lin (1974) made an evaluation of the autocorrelation and covariance methods of linear prediction by using synthetic vowels. In their study, vowels in the 'h-d' context were synthesized by a simulated formant synthesizer, and the two linear prediction methods were applied to analyze those synthetic vowels. As analysis conditions in this case, the sampling frequency was 10 kHz and the number of coefficients was 12. The results of their study are shown in Figures 5 and 6. Figure 5 shows the estimation error (in Hz) of the first three formant frequencies for both methods applied pitch-synchronously and pitch-asynchronously. For the pitch-synchronous case, the window length coincided with the segment position between the two pitch pulses. For the pitch-asynchronous case, the window length of 24 ms was arbitrarily chosen on the speech waves. The results indicate that the pitch-synchronous covariance method gives better accuracy than the others. In the pitch-asynchronous case, when the window length becomes greater than one and a half pitch period, the two methods give similar accuracy. The pitch-synchronous autocorrelation method resulted in the worst accuracy. This is more so in estimating formant bandwidths as shown in Figure 6.

For natural utterances, it is anticipated that the accuracy of estimating formant frequencies and bandwidths becomes worse

than for the synthetic sounds. Especially, it is anticipated that the result of the pitch-synchronous case will become worse, because the condition at the glottis varies during one pitch period for natural utterances, whereas the glottal condition for this particular synthesizer was constant. When the glottal condition varies during a chosen analysis segment, the resulting formant frequencies will probably be the average of the instantaneous formant frequencies. The result obtained by Chandra and Lin (1974) indicate that the pitch-synchronous covariance method gives more accurate estimates of formant frequencies and their bandwidths than the pitch-asynchronous autocorrelation method. Although the estimation accuracy of the formant bandwidths is not well known, it is known that the bandwidth estimates are sometimes too narrow or too broad. If the bandwidth information is needed, it has to be carefully checked against the direct Fourier transform of the corresponding sampled speech.

Problems in formant estimation

Since the estimation of formant frequencies is made from the envelope estimation of speech spectra, the accuracy of estimation is highly dependent on harmonic density. The more sparse the harmonic density becomes as pitch goes up, the more difficult the estimation of formant frequencies becomes. This is a rather inherent problem in the estimation of vocal tract resonances from given speech waves, irrespective of method. In many cases, the linear prediction method works well for speech sounds with fundamental frequencies of up to approximately 250 Hz. For female speakers and children with fundamental frequencies higher than 250 Hz, difficult cases of formant estimation are frequently observed. Formant estimation becomes impossible as the pitch becomes extremely high, in which case harmonics are picked up as spectral peaks.

In case the exact vocal tract resonances need to be known, some other methods may have to be used. One approach to this is to use external excitation with a low fundamental frequency such as an artificial larynx buzzer. One such example is shown in Figure 7 (a). This example is a female vowel /a/ with a fundamental frequency of 250 Hz. The linear prediction spectral envelope has one broad peak in the low frequency region instead of the first two formant frequencies. The peak-picking method de-

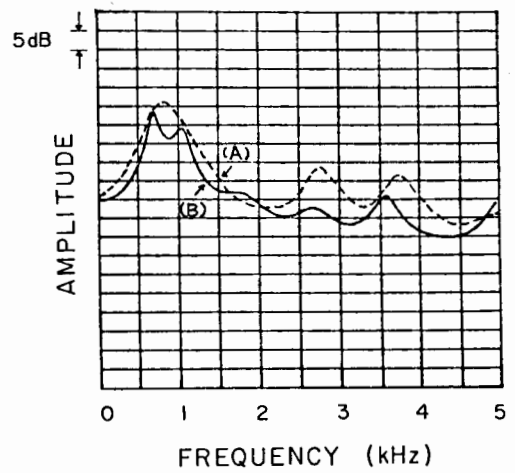


Figure 7. An example of difficult case of formant estimation. (a) Linear prediction spectral envelope for the vowel /a/ by a female speaker with a fundamental frequency of 250 Hz (sampling frequency 10kHz; number of coefficients 12; window size 25.6ms with a Hamming window and +6dB/octave preemphasis). (b) Linear prediction spectral envelope for the vowel /a/ by the same speaker excited by an external buzzer with a fundamental frequency of 80Hz (analysis conditions are the same as in (a)).

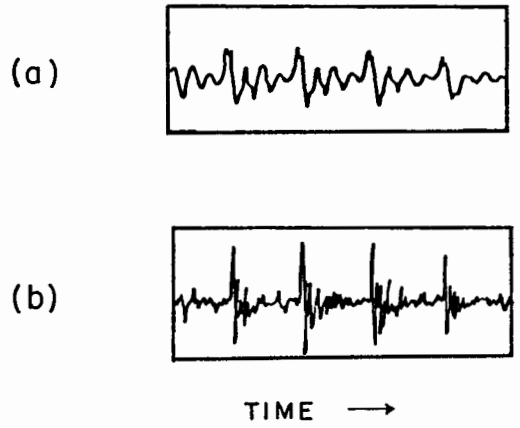


Figure 8. (a) Speech waves; (b) the residual signal after linear prediction analysis.

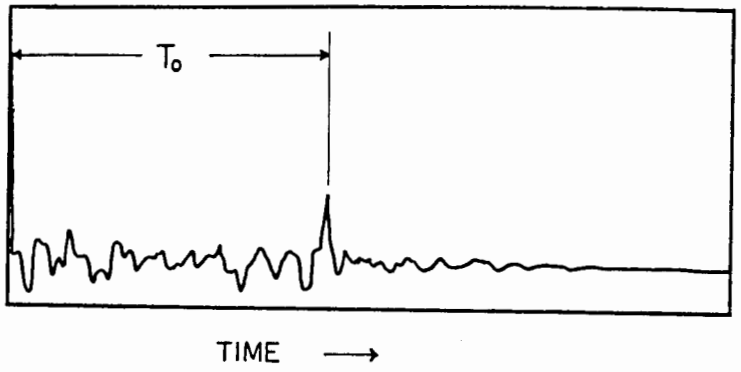
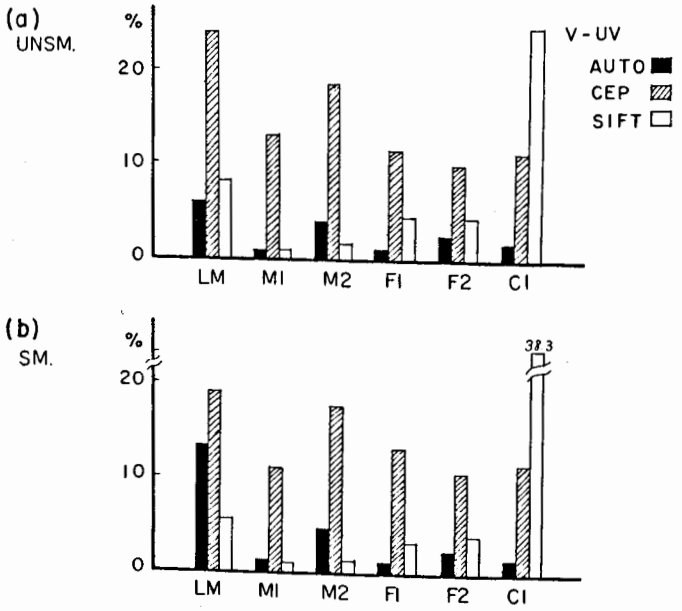


Figure 9. Autocorrelation function of the residual signal in Figure 8.



Figur 10. Voiced-to-unvoiced errors for three pitch detection methods: (a) unsmoothed; (b) smoothed. (LM:low-pitched male; M1,M2: males; F1,F2: females; C1: child). The ordinate shows the percentage error rate against total number of voiced intervals.

finitely fails to detect two peaks for F_1 and F_2. Instead it will detect the broad peak as the first formant frequency.

The root-solving method will give two roots to approximate the broad peak. It has not been ascertained, however, that the two roots obtained by the root-solving method for such cases as above correspond accurately to the first two formant frequencies. For the above case, the use of a commercial artificial larynx buzzer with a low fundamental frequency gives a good resolution for the formant frequencies as shown in Figure 7 (b), which is for the same vowel and the same speaker as in Figure 7 (a). In this case, the buzzer had undesirable sharp peaks in its own frequency characteristics. The monotonous frequency characteristics of a buzzer are desirable for this purpose.

Fundamental frequency estimation

In inverse filtering in the linear prediction method, most of the vocal tract characteristics are filtered out into the predictor coefficients. The residual signal, the output of the inverse filter, still contains the information on the excitation source. A typical residual signal is shown in Figure 8. It is seen that large errors synchronous with pitch periods occur. A typical approach to computing the periodicity from this kind of waveform is to compute the autocorrelation function as shown in Figure 9. Two conspicuous spikes are found in the autocorrelation function, one at the origin and one at a distance of one pitch period from the origin. The fundamental frequency is then given by the reciprocal of the pitch period.

Problems in fundamental frequency estimation

It has been shown that the linear prediction method is quite efficient and effective for estimating the formant frequencies. However, how accurate and reliable the extraction of fundamental frequency is is an intriguing question, since there are many other techniques for estimating the fundamental frequency. Rabiner et al. (1976), in their study of the comparative performance of several pitch detection algorithms, point out the following major problems in detecting the fundamental frequency: (1) glottal excitation is not perfectly periodic; (2) defining the exact beginning and end of each period is difficult; (3) the distinction between unvoiced portions and low level voiced portions is difficult; (4) there is an interaction between the vocal tract and the glottal excitation.

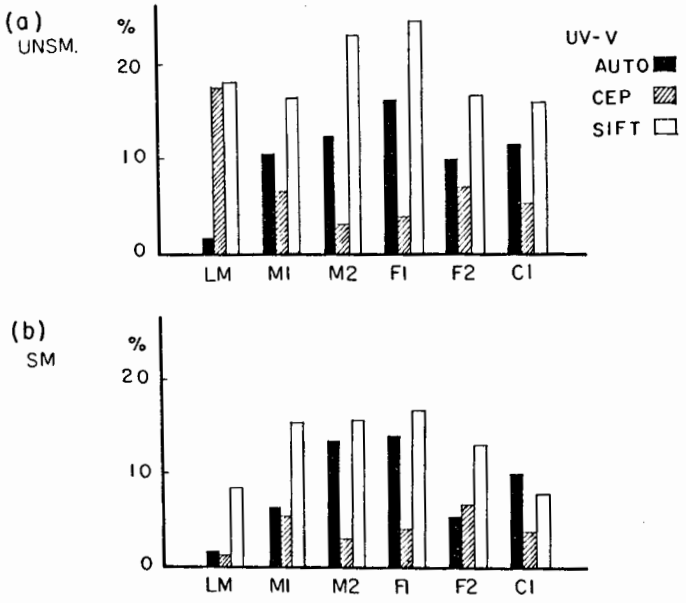


Figure 11. Unvoiced-to-voiced error for three pitch detection methods: (a) unsmoothed; (b) smoothed. (LM: low-pitched male; M1,M2: males; F1,F2: females; C1: child). The ordinate shows the percentage error rate against total number of unvoiced intervals.

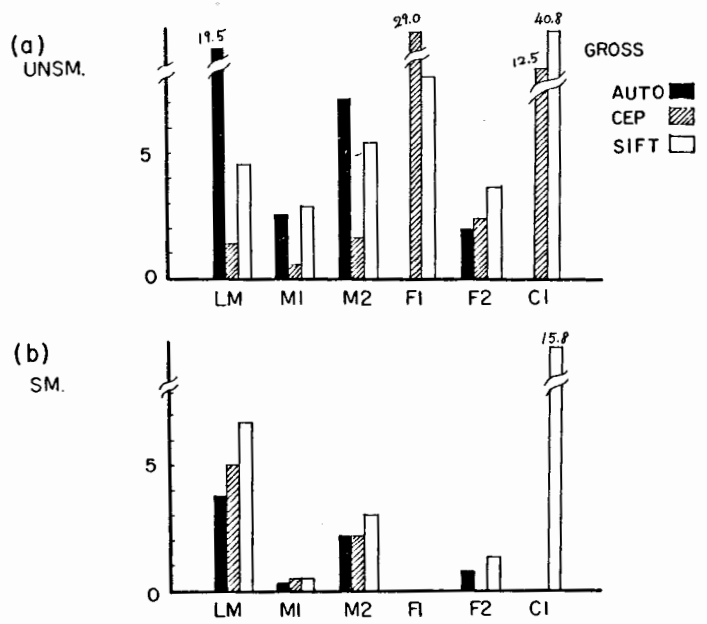


Figure 12. Gross errors for three pitch detection methods: (a) unsmoothed; (b) smoothed. (LM: low-pitched male; M1,M2: males; F1,F2: females; C1: child). The ordinate shows the average number of samples.

The above problems are intrinsic in any of the pitch detection methods. However, evaluation of several pitch detection methods indicates some differences in their performance.

Accuracy in fundamental frequency estimation

Let us take the following pitch detection methods from the study by Rabiner et al. (1975): (1) autocorrelation method with clipping (time domain method); (2) cepstrum method (frequency domain method); and (3) linear prediction 'SIFT'[1] method (time-frequency method). The types of errors can be categorized into (a) voiced-to-unvoiced error, (b) unvoiced-to-voiced error, (c) gross error in which the error in detecting the pitch period is greater than a certain threshold; and (d) fine error in which the error in detected pitch period is less than the threshold.

The above three methods were tested against six speakers (3 males, 2 females, and a child) by using four monosyllabic nonsense words and four sentences. The analysis results were compared with the standard pitch contours which were carefully measured by using a semi-automatic pitch detector. The results for the first three types of errors are shown in Figures 10, 11, and 12. The results are shown both for unsmoothed (raw data) and smoothed cases. In the smoothed case a nonlinear smoothing technique was applied to the raw data (Rabiner et al., 1975). It is seen that the nonlinear smoothing generally improves the accuracy: particularly, the gross errors are substantially improved. It is also seen that all three methods are somewhat speaker dependent. For the voiced-to-unvoiced errors, the error rate of the cepstrum method is much higher than the others except for the child speaker. For the unvoiced-to-voiced errors, on the other hand, the error rate of the cepstrum method is better than the others except for one of the female speakers for the smoothed case. In overall performance evaluation, there seems to be not much difference between the performance of the autocorrelation and linear prediction methods, except that the linear prediction method resulted in an exceedingly poor performance for the child speaker for the unvoiced-to-voiced and gross errors.

Other related topics

The filter box in the linear prediction model in Figure 1 contains the contribution from the glottal characteristics and the radiation effect at the lips as well as the vocal tract

1) Simplified Inverse Filter Tracking

characteristics. Since the model assumes a linear system, those factors can be separated and changed in order as shown in Figure 13. If the glottal and radiation characteristics can be eliminated by a proper preprocessing of the speech, the true vocal tract characteristics can be obtained by the linear prediction method. One of the important features of the linear prediction method is that in computing the prediction coefficients, another parameter which is called "reflection coefficient" (or "k-parameter", or "PARCOR coefficient") is obtained. A set of reflection coefficients obtained for a given speech segment gives an acoustic tube shape which has a frequency characteristic identical to the vocal tract characteristics extracted from this speech segment. In this case, the acoustic tube is represented by a concatenation of cylindrical sections of different cross-sectional areas. A reflection coefficient is defined at the boundary between two neighboring sections. Consequently, if the analysis conditions are properly chosen after preprocessing sampled speech to eliminate the glottal and radiation characteristics, the acoustic tube representation thus obtained is expected to be a good approximation to the vocal tract area function which denotes the cross-sectional areas along the vocal tract from the glottis to the lips (Wakita, 1973, 1979).

Another interesting topic is the use of the linear prediction parameters for speech synthesis. The synthesizer could be the synthesis part of the linear prediction analysis-synthesis telephony (see Markel and Gray 1976; Wakita 1976). Since the formant frequencies and bandwidths constitute the roots of the inverse filter transfer function, they can be related to the filter coefficients. The reflection coefficients, which give an acoustic tube representation of the vocal tract, are also related to the filter coefficients in the mathematical formulation of linear prediction. Thus, those parameters mentioned above are interchangeable for each other, and any of these parameters can be used for the linear prediction synthesizer.

Application examples

The linear prediction method has mainly been used in the area of analysis-synthesis telephony. The method is particularly effective for low bit-rate speech coding. However, the technique is equally useful for acoustical analysis of speech. In concluding this tutorial paper, several examples taken from the author's past studies will be given below.

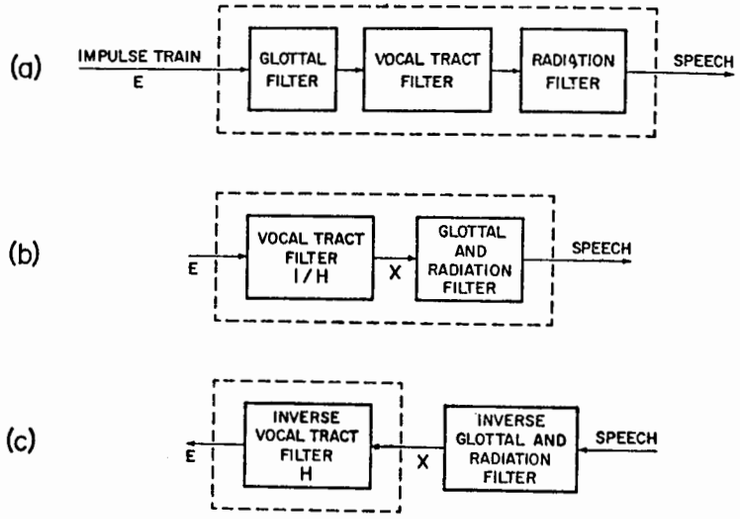


Figure 13. Block diagrams to obtain the vocal tract characteristics by eliminating the glottal and radiation effects.

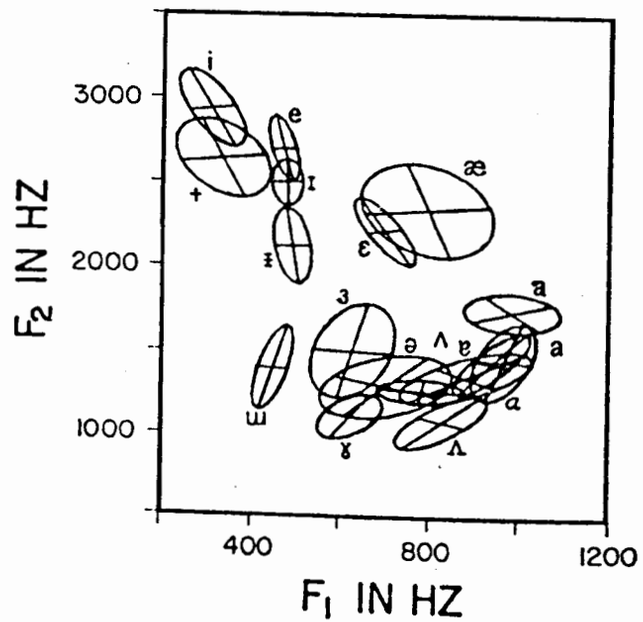


Figure 14. F_1-F_2 distribution of 17 unrounded vowel types produced by a female speaker. Ellipses represent two standard deviations.

Example 1 (Broad and Wakita, 1978).

Figure 14 shows the F_1-F_2 distribution for 17 unrounded vowel types produced by a female phonetician in order to study the variability of formant frequencies. In this study, 30 repetitions of 30 different isolated vowel utterances (900 in total) were analyzed by the linear prediction autocorrelation method (sampling frequency 10 kHz; number of coefficients 12; window size 25 ms with Hamming window and +6 dB/octave preemphasis) and formant frequencies were estimated by using the root-solving method. The analysis results were carefully inspected by displaying the results vowel by vowel on the display terminal. Approximately 5% of apparently wild data were excluded for further processing.

Example 2 (Wakita, 1977)

The example in Figure 15 shows the F_1-F_2 distribution of nine American English vowels spoken by 26 speakers (14 males and 12 females) in order to study the variability of formant frequencies among male and female speakers. Vowels were produced in the context of 'h-d' and the linear prediction autocorrelation method was applied to analyze the vowel portions (sampling frequency 10 kHz; number of coefficients 12; window size 25 ms with Hamming window and +6 dB/octave preemphasis). The formant frequencies were estimated by use of the root-solving method. The formant frequencies which were averaged over the most steady-state five frames were used to represent each vowel.

Example 3 (Kasuya and Wakita, 1979)

Figure 16 is an example in which the linear prediction area functions were used to automatically segment speech into vowel-like and nonvowel-like intervals. The linear prediction area functions, combined with the speech energy function (root mean square of sampled speech), give sufficient cues for the first stage of segmentation without obtaining spectral information which is more time consuming. In this case, the autocorrelation method was also used for analysis. The sampling frequency was 10 kHz, the number of coefficients 14, and window size 15 ms with a Hamming window and +6 dB/octave preemphasis. The relatively short analysis window length was used in this study for detecting the bursts of plosives, and the window shift was 12.8 ms.

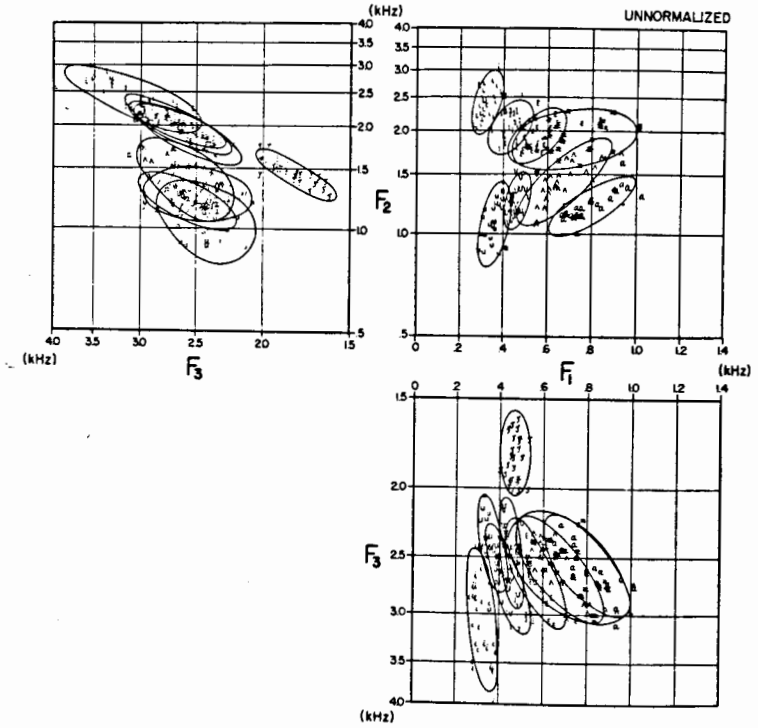


Figure 15. Distribution of formant frequencies projected onto the F_1-F_2, F_1-F_3, and F_2-F_3 planes for 26 speakers (14 males and 12 females). Ellipses represent two standard deviations.

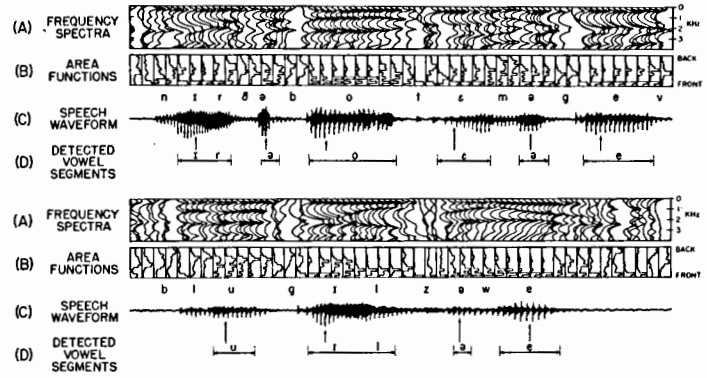


Figure 16. An example of segmenting the vowel-like intervals for the sentence " Near the boat, Emma gave blue-gills away."

Conclusion

The concept and evaluation of the linear prediction method were described in this paper. Because of its tutorial nature, the descriptions in some cases may be inadequate from the theoretical point of view. Readers interested in more advanced knowledge are encouraged to read the original papers or other materials listed in the references.

Acknowledgement

The author would like to thank Dr. P.-A. Benguerel, The Phonetics Laboratory, University of British Columbia, Canada, for his collaboration in investigating the use of an artificial larynx buzzer.

References

Atal, B. and M.R. Schroeder (1967): "Predictive coding of speech, Proc. 1967 Conf. Commun. and Process., 360-361.

Broad, D.J. and H. Wakita (1978): "A phonetic approach to automatic vowel recognition", in Bolc Speech communication with computers, 52-92, London: Macmillan.

Chandra, S. and W. Lin (1974): "Experimental comparison between stationary and nonstationary formulation of linear prediction applied to voiced speech analysis", IEEE Trans. ASSP-22, 403-415.

Itakura, F. and S. Saito (1966): "A statistical method for estimating speech spectrum", Technical Report 3107, Electrical Commun. Res. Lab., NTT.

Kasuya, H. and H. Wakita (1979): "An approach to segmenting speech into vowel- and nonvowel-like intervals", IEEE Trans. ASSP-27, 319-327.

Makhoul, J. (1975): "Linear prediction: a tutorial review", Proc. of IEEE vol. 63, 561-580.

Markel, J.D. and A.H. Gray (1976): Linear prediction of speech, New York: Springer.

McCandless, S.S. (1974): "An algorithm for automatic formant extraction using linear prediction spectra", IEEE Trans. ASSP-22, 135-141.

Rabiner, L.R., M.R. Sambur and C.E. Schmidt (1975): "Applications of a nonlinear smoothing algorithm to speech processing", IEEE Trans. ASSP-23, 552-557.

Rabiner, L.R., M.J. Cheng, A.E. Rosenberg and C.A. McGonegal (1976): "A comparative performance study of several pitch detection algorithms", IEEE Trans. ASSP-24, 399-418.

Wakita, H. (1973): "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms", IEEE Trans. AU-21, 417-427.

Wakita, H. (1976): "Instrumentation for the study of speech acoustics", in Lass (ed.) Contemporary issues in experimental phonetics, 3-40, New York: Academic Press.

Wakita, H. (1977): "Normalization of vowels by vocal-tract length and its application to vowel identification", IEEE Trans. ASSP-25, 183-192.

Wakita, H. (1979): "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: the state of the art", IEEE Trans. ASSP-27, 281-285.

DISCUSSION

Gunnar Fant, Wiktor Jassem and René Carré opened the discussion.

Gunnar Fant: I think that at the moment LPC analysis is more useful for communication engineering purposes, but it is certainly gaining importance in phonetic analysis: the fact that you can re-synthesize speech with rather good quality with LPC methods is a great advantage in synthesis, and LPC also makes it possible to manipulate e.g. fundamental frequency, independently of other parameters, which makes it well suited for prosodic investigations.

Formant frequencies and bandwidths describe the vocal filter, but what about the vocal source? In LPC analysis, it is treated as a constant function, more or less, but in the future we should pay more attention to the time dynamics of the source, to obtain valuable information for prosody studies. We should make dynamical matches not just to formants but also to source characteristics. (This we can do at present by carefully scrutinizing period after period of the signal, extracting presumed vocal source characteristics.) The fact that LPC is confined to an on/off, or voiced/voiceless, distinction creates some undesirable compensation effects: to compensate for a more steeply falling voice source spectrum, like we get e.g. in open syllables, the system will increase the bandwidths somewhat, which can give a consonantal effect.

Another critical problem is assessing formant frequencies with high pitched voices and in cases where Fo and F1 are close together, which is problematic in any kind of analysis.

Hisashi Wakita: mentioned a comprehensive LPC analysis of 900 vowels by a female speaker (30 vowels x 30 repetitions) where (50) unlikely analysis items were discarded by visual inspection of the vowels in F1-F2, and F1-F3 plots [see "Application Examples", Example 1 in Hisashi Wakita's paper], but admitted that we do not yet have valid data that tell us how accurately we can estimate formant frequencies, especially when Fo and F1, or two formants, are close together.

If we analyse a little more than one pitch period, using a very small time window and the covariance method we can, from the error signal, determine that point where the interaction between sub- and supraglottal systems is minimum (corresponding to the

closed glottis portion), and if the signal has been carefully recorded, directly from the microphone into the computer storage, so as to avoid phase distortion, we can fairly well recover the glottal wave shape from this portion.

Wiktor Jassem: What is the perspective for phonetics of these methods? First, there is the segmentation problem which can probably be solved, as suggested by professor Fant and others, by determining the maximum rate of change of the spectrum and of the time function. Secondly, there is the extraction of parameters: those extracted for automatic analysis need not be identical to those used by a human being. Thirdly, there is the problem of normalizing for individual speaker characteristics. The fourth problem is concerned with the identification of entities, which is an intricate one, because we do not know how many entities there are. The theory is that they should be sufficient to specify the output in such a way that synthesizing it we would get a normal native accent. The perceptual experiments needed to settle the question are not simple, because the adults' responses will be heavily influenced by phonemic considerations, and with very young children there will be great psychological problems. Fortunately, mathematical methods are developing that will allow us to determine, given a number of data, how many objects or entities we are dealing with. What I want to point out is that if we can get the computers to do phonetic transcriptions they will be better than transcriptions by a human being because they will be more objective.

René Carré: There are two kinds of work in speech analysis. One is the analysis of a small number of speech sounds. Formant frequencies are no problem, but to determine bandwidths we need to consider pre-emphasis, the order of the predictors, the analysis window, and the magnitude of the prediction error. All these operations take time, and such a procedure cannot be adopted in the other kind of study, of a large corpus, where a (semi-)automatic procedure has to be set up. It seems that in that case the procedure must be normalized. Is the autocorrelation method accurate enough for bandwidth measurements? Must we change (automatically or not) the order of the predictor to adapt the system to the speech sound under analysis, e.g. to nasalized vowels? What sampling rate shall we choose? How many frames should be analyzed? And so on. Finally, among the set of pole values we have to choose (automatically or not) the right formants.

Hisashi Wakita: The RMS-function is generally not sufficient to segment a chain into vowel-like and non-vowel-like sounds. But from the pseudo vocal tract area function, generated by the LPC analysis, we can calculate the ratio of the volume of the back (pharyngeal) cavity to the total volume of the vocal tract and this will generally tell us whether a segment is vowel-like or not. It will detect nasal consonants which is difficult to do from the waveform: LPC does not assume any nasal tract, but does produce a sort of equivalent acoustic tube representation, and nasal segments are fairly well detected from the back-to-total ratio of that tube.

We have also worked on the elimination of inter-speaker variability, which is of interest not just to automatic speech recognition, but also in acoustic phonetic studies of e.g. the vowel systems of languages. With LPC we can estimate the vocal tract length for each speaker and each vowel category (tract length is not constant over different vowel qualities), and then normalize to a certain length, e.g. 17 cm, a normalization which reduces the overlap in F1-F2, and F1-F3 plots and results in compact vowel distributions.

Adrian Fourcin: The LPC system represents the complexities of the vocal tract and its excitation by an exceedingly simple model: a vocal tract with no side-branches and a sharp impulse for an excitation, and yet it produces speech of very high quality. When we synthesize we have to pay attention to the zeros introduced by nasality, and the time dependence of the excitation function is also apparent if we have a standard model of the vocal tract. Is there something that we can learn from this with regard to how we hear speech?

If we knew when the point of excitation occurred and for how long a time the glottis is closed, to what extent would you be able then to improve the phonetic utility of the LPC analysis?

Hisashi Wakita: The ear is insensitive to spectral zeros, and a model which has poles and zeros in it (which is much more complicated computationally) does not perceptibly improve the quality of the speech. I have run an experiment, where various musical instruments as well as speech were passed through an artificially generated pole-zero system, and it turned out that the ear was insensitive to dips in the spectrum as large as 35 dB

(a fact which explains why HiFi loudspeakers may have even very sharp dips).

If we can determine that segment of speech where the glottis is closed, i.e. the force-free oscillations, we can apply the covariance method, which assumes that the speech waves can be approximated by a combination of damped sinusoids, and thus compute the exact vocal tract characteristics.

Gunnar Fant: A reply to Dr. Fourcin is that LPC speech sounds good because it resembles natural speech, although its source and transfer functions do not resemble those of real speech. The source function is stylized, but then there is a compensation in terms of the transfer function chosen to get the overall result correct (something which invalidates the data we get on formant frequencies and bandwidths).

Another characteristic of LPC analysis is that all the losses are concentrated at the glottal end of the system. How much does that invalidate the bandwidth data?

Hisashi Wakita: It is true that the LPC method approximates the spectral envelope, without any regard to formant frequencies and bandwidths. All the energy losses are lumped into one single resistance at the glottis end. By means of this single resistance we represent all the bandwidths of the spectrum. If we want to relate it to a particular speech production model, in terms of formant frequencies and bandwidths, it is quite useless, I think, so either we have to build more realistic models, both production and inverse transform models, or we can try to relate the simple LPC model to a more realistic, complicated model.

John Clark: There seems to be no great difference in the intelligibility levels quoted in the recent literature for predictor coded and formant coded speech. For formant coded speech, some of its phonetic weakness appears (when tested with CV-nonsense syllables) in the fricatives. Is this also the case for predictor coded speech, and what sort of evaluation have you done of the perceptual weaknesses of the system as a means of synthesizing speech?

Hisashi Wakita: Normally, with the LPC analysis-synthesis we use the extracted coefficients as they are, but we replace the residual signal with a pulse train which makes the voiced/unvoiced decision very critical, and missing just one frame can be perceptible. We can, however, restore the original signal by using the residual signal for excitation. For phonetic evaluation purposes I think we have to choose the excitation source carefully, - maybe not the residual signal itself, but one with which we do not loose too much information about the source.

SYMPOSIUM NO. 1: PHONETIC UNIVERSALS IN PHONOLOGICAL SYSTEMS AND THEIR EXPLANATION

(see vol. II, p. 5-59)

Moderator: John J. Ohala

Panelists: Thomas V. Gamkrelidze, André-Georges Haudricourt, Robert K. Herbert, Jean-Marie Hombert, Björn Lindblom, Kenneth N. Stevens, and Kenneth L. Pike

Chairperson: Bertil Malmberg

JOHN J. OHALA's INTRODUCTION

Phonetic universals is such a large subject that the members of this symposium despaired of being able, in the short time allotted, to give adequate consideration to any of the general aspects of the theory or practice of the field or to solve any of its "great problems". It was decided, therefore, that the moderator would make a few brief general comments about some of these larger issues, more or less "for the record", but that most of the time of the symposium be devoted to the discussion of one very specific problem in the area of phonetic universals.

General Problems and Issues in Phonetic and Phonological Universals

(In this report I will use the shorter phrase 'phonological universals' for the longer, somewhat unwieldy expression 'phonetic universals in phonological systems', the official topic for this symposium.)

1. Before beginning this discussion, we should define what we mean by *phonological universals*. As this term has come to be used, it means *systematic patternings of speech sounds cross-linguistically*. This definition does not require that the pattern be manifested in every human language, merely that it have sufficient incidence in the languages of the world such that its occurrence could not be attributed to chance. It is assumed, though, that all languages, indeed, all human speakers, are potentially subject to whatever "forces" create these patterns, but an overt manifestation of these forces may or may not occur and if it does occur, may take different forms. For example, to consider a case discussed extensively by Professor Gamkrelidze, it is presumably the same universal factors which are responsible for the asymmetrical gap in the voiced velar stop position (/g/)

in the segment inventories of Dutch, Czech, and Thai, as are responsible for the disproportionately low incidence of /g/ in the lexicon or in running speech of many languages. Likewise, whatever causes the asymmetrical absence of /p/ in Arabic, Nkom, and Chuave, is also responsible for the limited distribution of /p/ in Japanese, i.e., it only appears intervocalically and as a geminate.

2. The concern with phonological universals in our field has both theoretical and practical consequences. Some 100 years ago our intellectual forefathers, Ellis, Sweet, Passy, Lepsius, Jespersen, and others, provided us, in the phonetic alphabet and the descriptive anatomical and physiological terms accompanying it, the equivalent of the Linnean system of classification in biology or Mendeleev's periodic table of the elements in chemistry. Today, I believe it safe to say that we have reached the stage equivalent to that which Bohr's model of the atom represented in physics and chemistry. We have a framework within which to observe, to describe, and to establish natural classes of phonetic and phonological entities and processes in all human languages. We are also able, with obvious limitations, to predict and explain the behavior of speech sounds. Commendably, in many cases, these explanations are based on empirically-supported models of parts of the speech communication process. Although it is obviously the case that as we deepen our understanding of some of the basic physical, physiological, and psychological mechanisms serving speech, we also are better able to explain many phonological universals; it is also true that in many cases ***it is our observation of phonological universals which leads to a greater understanding of speech mechanisms***. The literature in phonological universals is even now causing us to critically re-examine some of the most fundamental concepts in phonetic and phonological theory, for example, the notions of 'segment', of 'distinctiveness', etc., and to explore in considerable detail in the laboratory basic acoustic, aerodynamic, and auditory mechanisms in speech.

In the practical realm phonological universals can aid us in the analysis and understanding of the phonologies of individual languages: they tell us what to look for and they help us to choose alternative scenarios for the history of sound changes in the language. I personally believe that phonological universals

can also aid us in such cases of ***applied phonology*** as speech synthesis, automatic speech recognition, speech pathology, speech therapy, and language teaching. It must be said, however, that at present there has been very little penetration of universals in these areas.

3. Phonological universals are found in many different forms, e.g., segment inventories, segmental sequential constraints ("phonotactics"), allophonic variation, sound change, morphophonemic variation, dialect variation, patterns of sound substitution by first and second language learners, frequency of occurrence of sounds in the lexicon and in connected speech, conventional and esthetic use of speech sounds in onomatopoeia, poetry, jokes, singing, etc. Can we bring all of these disparate phenomena under one theoretical umbrella, using one of these as the base or primitive from which the others may be derived, or, possibly, deriving them from some separate principle external to all of them?

4. Another general issue concerns the problem of how to obtain a truly representative sample of sound patterns from a variety of languages such that the sample is not biassed by including too many or too few languages having certain genetic, typological, or geographical linkages. The many pitfalls of attempting a quantification of phonological data from large samples has been discussed previously, including such concerns as how one differentiates a language from a dialect, whether one should look at the behavior of phones or phonemes and if phonemes, whose conception of the phoneme, etc? The fact is, most works on phonological universals ignore this issue and seem to rely on the investigator's intuitive "feel" for what constitutes a proper sample. Is there any way to make this process objective? How can we create an unbiassed sample; how large should it be?; what criteria should we apply in admitting a language to the sample? Once we have the supposedly unbiassed sample, what type of statistical analysis should we apply to it in our attempts to prove or disprove universal tendencies?

My own solution to this problem, a solution which has parallels in other scientific disciplines, is to make sure that any posited universal is supported both ***inductively*** -- that is with lots of examples (and few counterexamples) - and ***deductively*** --

that is, by what we know to be the underlying operating principles of speech production and perception.

5. A related issue is whether or not some of the claims made about phonological universals may be distorted by observer bias, i.e., be self-fulfilling prophecies. It has been claimed, for example, that all languages code speech in terms of phonemes. But I know of no universally-accepted algorithm which discovers phonemes. And if there were, do we now have any evidence that phonemes and all the properties attributed to them, have psychological and/or physical reality?

A very clear example of the perils of observer bias surrounds claims about universals of syllable structures. It has been claimed that within a syllable, one should not find a transition from voiced to voiceless to voiced. Upon being presented with an apparent counterexample such as [itv], the claimant would protest that there is a syllable boundary between the [t] and [v]! The potential for similar circularity enters into any claim which contains terms that cannot be objectively defined. And this, unfortunately, is true of a very large number of terms used in phonetics and phonology, including terms such as consonant, vowel, segment, syllable, sonority, strength, lenition, etc.

Would we find a different set of universals if we adopted the parallel, hierarchic system such as Professor Pike advocates? Would we have a different, more interesting set of universals if we included in the description of sounds, as Professor Stevens proposes, the sensory information each sound gives rise to?

A Specific Problem in Phonological Universals

The problem selected for special attention during this symposium is by no means a small one and it is doubtful that it will be solved very quickly, certainly not in the short time allotted us. Nevertheless, it is a problem that intersects with the particular interests of most members of the symposium and is a matter to which many members of the audience can contribute. The problem is stated in a deliberately provocative way in order to stimulate discussion.

The notion of a vowel "space" has been used in phonetics for about 2 centuries but it is only recent evidence which points to this space having acoustic-auditory correlates. The research of Lindblom and his colleagues suggests that the placement of vowels in this space in various languages is dictated by the principle of maximal perceptual difference, i.e., that however many vowels there are in the system, they tend to arrange themselves in the available space in such a way as to maximize their distance from each other. This principle seems to adequately predict the arrangement of systems with approximately 7 or 8 vowels. It would be most satisfying if we could apply the same principles to predict the arrangement of consonants, i.e., posit an acoustic-auditory space and show how the consonants position themselves so as to maximize the inter-consonantal distance. Were we to attempt this, we should undoubtedly reach the patently false prediction that a 7 consonant system should include something like the following set:

ɗ, k', ts, ɬ, m, r, ɜ,

Languages which do have few consonants, such as the Polynesian languages, do not have such an exotic consonant inventory. In fact, the languages which do possess the above set (or close to it), such as Zulu, also have a great many other consonants of each type, i.e., ejectives, clicks, affricates, etc. Rather than maximum differentiation of the entities in the consonant space, we seem to find something approximating the principle which would be characterized as "maximum utilization of the available distinctive features". This has the result that many of the consonants are, in fact, perceptually quite close -- differing by a minimum, not a maximum number of distinctive features.

Does this mean that consonant inventories are structured according to different principles from those which apply to vowel inventories? Could it mean that the "spaces" both consonants and vowels range in, are limited by the auditory features (= parameters) recognized by the particular language? Or does it mean that we are asking our questions about segment inventories in the wrong way?

COMMENTS FROM THE PANELISTS

K.N. Stevens: In an acoustic representation of connected speech we find certain regions where there are rapid (10-30 msec) changes in a number of acoustic parameters, e.g., amplitude, periodicity, and spectrum. A hypothesis that has emerged from our and Chistovich's research, is that the attention of the listener is drawn

to these regions, more so than to other regions where changes are less rapid. These regions are, first of all, markers of consonants, but additional information can also be packaged in them along several orthogonal dimensions. We believe languages therefore tend to "select" a consonant inventory that uses up most of these dimensions. These primary dimensions are: [± voice] (presence/absence of periodicity), [± nasal] (presence/absence of low-frequency murmur), [± continuant] (unbroken/interrupted sound), [± grave] (low-/high-frequency tilt to the spectrum), [± compact] (energy spread out/concentrated). After processing the information in these regions of rapid change (= high rate of information transfer), the listener's attention may focus on the remaining regions and here lie the cues for such dimensions as palatalization, pharyngealization, clicks, etc. It logically follows that the learning of (or introduction of) such distinctions will *follow* the learning of distinctions coded in the regions to which primary attention is directed.

B. Lindblom: We have recently followed up and improved on our early work on predicting vowel inventories and I think the research strategy we have used could be applied to consonant inventories, too. Briefly, our procedure is to 1) specify a physiological model of the vocal tract and use it to define 2) the range of humanly possible vowels and from this derive 3) the (universal) human acoustic vowel space, a continuum, and, finally, 4) to employ an auditory model to define a perceptual space to accommodate a specified number of vowels. The last step consists of convolving an input power spectrum (of a given vowel) with an auditory filter derived from masking data, thus yielding a hypothetical auditory excitation pattern. We assume that, other things being equal, the probability of any two vowels being confused, that is, their perceptual closeness, will be related to the overlap area enclosed by their excitation patterns. We believe vowel systems evolve so as to make vowel identification efficient and this is done by making perceptual differences between vowels (quantified as mentioned above) maximally or, perhaps, sufficiently large. This new measure of perceptual distance yields much more reasonable predictions about vowel placement; in particular, it eliminates the excessive number of high central vowels that plagued previous models.

A preliminary typological study of diphthongs shows that [a^i] and [a^u] are the most favored. This result is compatible with the new properties of our model's perceptual space and provides evidence for a principle of perceptual differentiation applying not only paradigmatically, but also sequentially. Consonant inventories can be studied within a paradigm such as this.

K. Pike: My own approach to phonetic analysis is a bit different from that of most of my fellow panelists. Although I have often been helped by acousticians when I have brought my phonetic problems to them, I would rather argue that the reductionism, so necessary in the laboratory, is detrimental to linguistic analysis in the field. I can illustrate this with an examination of a short poem by E.E. Cummings. [Text and detailed commentary omitted.] Although one can point out puns, details of orthography, prosody, and even cultural allusions which contribute to the overall effect, the poem, like language, functions as a whole. I am encouraged by the enlarged scope of phonological inquiry demonstrated at this congress, e.g., the work on syllables. The study of vowel spaces should also be enlarged to include what I call 'pharynx space' (changes in vowel quality by modifications of pharyngeal width and larynx height) and by taking into consideration the psychological reality of vowel structure.

J.-M. Hombert: A surprising number of people I have met at this congress are quite skeptical about the existence of phonological universals. Although one can cite countless examples of cross-language similarities in sound inventories, sound changes, and phonological processes, there are, of course, always counterexamples to almost any generalization one might make. Perhaps the answer to this is to pay more attention to the diachronic aspect of universals: the counterexamples may just be unstable transitional states between more natural states. Moreover, it is often possible to find that certain cited counterexamples cease to be so if one looks into the details more closely, e.g., in cases of tonal development from obstruents, a voiced stop giving rise to a high tone runs counter to the usual patterns, but if it was found that the voiced stop had first become an implosive, an expected development, then the case is no longer a counterexample.

Concerning the sampling problem, mentioned by the moderator, it is particularly acute in the case of perceptual data. This can be solved if we start discovering ways to take our laboratories in-

to the field and thereby gather perceptual data from a wide variety of languages.

R. Herbert: A consideration of the factors constraining the introduction into a consonant inventory of complex sound types, e.g., affricates, pre- and post-aspirated consonants, and especially prenasalized consonants, may provide insight into the constraints on consonant inventories as a whole. Obviously, the parts of such complex segments must be sufficiently different from each other so that they may both be perceptually salient within the time span of a single segment, e.g., the nasal/oral distinction used in prenasalized stops. It must also be possible to articulate the parts within this same time span. Thus there are limits on the number of components in single segments: usually 2, but 3 in the case of pre-nasalized affricates, and rarely more. Most such complex sounds involve at least quasi-homorganic components, and thus nasal and stop combinations are frequently encountered but lateral and stop combinations less so since laterals, unlike nasals, have limited capacity for homorganicity. We might also speculate that the relative ordering of the components in complex segments is governed by the same factors that determine optimal syllable codas: the first element is generally the more common syllable coda, it being understood that optimal syllable codas are drawn first from the opposite ends of the sonority hierarchy, e.g., glides, nasals, [ʔ], and voiceless stops, before involving segment types from the middle, e.g., laterals, voiced stops, fricatives.

A.-G. Haüdricourt: The search for phonological universals seems to me to be like the quest for the philosopher's stone. As for phonetic changes, it is more profitable to look at the conditions for the appearance of the phenomena rather than for their existence. Language is a social phenomenon and one of its main functions, communication, causes the development of new phonemes. Sindhi provides an example: its whole series of voiced stops, when long, has become preglottalized in order to remain distinctive. Language also has a socio-ethnic function and so preglottalization may appear without any phonological conditioning, as happens in Vietnamese and the Henan dialect of Chinese. In these cases, one or two preglottalized consonants are sufficient for the social function and it is normal that they should be the easiest to articulate (ɓ, ɗ). Likewise, preglottalized consonants can disappear for a variety of reasons. The loss of these sounds in Vietnamese was in part due to the presence of tones (which made the voicing superfluous) but has also been aided by the sociolinguistic environment in, e.g., Saigon. These facts are outside the domain of instrumental phonetics.

T.V. Gamkrelidze: I believe an understanding of the principles governing the structure of consonant and vowel inventories will come from typological phonology and experimental phonetics. An important task for typological phonology today is the establishment of constraints or relations of markedness or dominance between certain bundles of co-occurring features. For example, as detailed in the printed version of my paper, in the subsystem of stops and fricatives, $[^{+voice}_{+labial}]$ is dominant (unmarked) with respect to the co-occurring features $[^{+voice}_{+velar}]$. Thus, among voiced stops, /b/ is dominant, /g/ is recessive. Also, among voiceless stops, /k/ is dominant, /p/ is recessive. These relations stem from the specific acoustic and articulatory properties of the features involved. In the examples mentioned, the volume of the air chambers plays a part. Gaps in the paradigmatic system of obstruents will generally reflect these dominance/recessiveness relations. These relations can therefore help us to better understand sound change and to do language reconstruction more realistically. In light of this, the classical reconstruction of the Indo-European occlusive phonemes appears to be linguistically improbable in that (among other things) it assumes the series with the missing labial were voiced stops. Reinterpreting this series as ejectives brings the IE obstruent system into full conformity with typological studies.

J.J. Ohala: I would speculate that a universal vowel and consonant space does not exist. Each language "chooses" some restricted set of features or dimensions for these spaces. It is common knowledge, for example, that a native speaker of one language is 'deaf' to certain features used in other languages. It is true that the Lindblom model does have a remarkable degree of success in predicting the structure of systems with a small number of vowels. But it is significant that it breaks down when a large number of vowels are involved, very likely because one or two dimensions other than those used in the model are also involved, e.g., vowel duration, diphthongization, voice quality. It could be that vowel spaces, unlike consonant spaces, have rather few possible dimensions and that most languages make some use of the most salient dimensions (those based

on spectral shape). In consonant systems, it is well known that there are more possible dimensions to choose from and so the discrepancy between reality and the predictions of a maximum-perceptual-distance model are more evident. Thus, the differences between vowel and consonant systems in this respect are only apparent. What is more remarkable --to me, at least -- is the highly symmetric nature of consonant proliferation. The mechanism of proliferation is reasonably clear, e.g., stop plus [ʔ] yields a glottalized series of stops or ejectives, but why should proliferation almost always yield a whole new row or column of such consonants?

DISCUSSION

K.N. Stevens: It is true, as Professor Gamkrelidze notes, that aerodynamic factors contribute to the asymmetries in obstruent systems, but auditory factors are important, too. The noise or burst of a voiceless velar will give a very clear indication of compactness -- more so than a voiced velar, whereas a voiced labial will reveal the feature [+grave] better than a voiceless labial.

J. Ohala and K. Stevens discussed the need, in the search for the most salient auditory dimensions, of finding the perceptual cues for such striking sounds as ejectives.

K. Pike and J. Ohala mentioned specific instances of vowel and consonant systems utilizing voice quality as a distinctive dimension, e.g., certain languages of Nepal, various Nilotic languages, Korean, Javanese, Cambodian, Gujarati.

B. Lindblom: It is possible, in principle, to include other dimensions in the vowel space, but it is better at this stage of research to make our models precise and quantitative. At present then, it is better to restrict the investigation to spectrally-based dimensions. I agree with Ohala that listeners react to vowel stimuli in language-specific ways. In fact, some of our own research shows that Swedish listeners put more subjective distance between the vowels in the crowded front region of the Swedish vowel space than would have been predicted by our model's spectrum-based metric. But let us not be too hasty in discarding the notion of a universal vowel space. After all, this may be what the child brings to the language-learning task.

J. Ohala: I concede that I overstated my position. There undoubtedly is a universal vowel space and each language chooses a subspace within it. No doubt there is some order according to which features are chosen first.

T.V. Gamkrelidze: The greater proliferation of consonants as opposed to vowels is due to the greater number of possible dimensions in consonant systems. In theory, of course, an infinite number of vowels could be produced, but practically the number is small due to auditory and articulatory constraints.

A. Haudricourt: (In response to a question from J.-M. Hombert) The search for phonological invariants and for culture-specific phenomena is not incompatible, but they are two different problems. First we must investigate the *function* of language and only then look at its phonetic realization.

B. Lindblom: Given the well known *discreteness* of language, it might be asked why, in our model, we start with a *continuous* vowel space. The answer is that we do not yet have a theory that predicts that language should have discrete units such as distinctive features. The theory of distinctive features we do have is based on induction. I think the discreteness has to be deduced or derived as a consequence of more fundamental principles. Even so, a totally discrete model will still not explain why, in languages with few vowel contrasts, the extreme corner vowels tend to be phonetically less extreme (as noted by Crothers).

(To Prof. Stevens:) The quantal phenomena you find in the articulatory-to-acoustic transformation cannot be the only source of phonological discreteness. Surely, memory mechanisms must be involved as well (cf. the work of G. Miller and I. Pollack on elementary auditory displays).

K. N. Stevens: I agree with all of your points. I would just say that in the vowel space there are some regions which are more stable (or discrete) than others in that a wide range of articulations would give rise to the same acoustic signal. So the vowels will be within these regions, the exact location determined by factors such as your model incorporates. It is possible, too, that the whole space may shift in one direction or another due to different so-called 'basis of articulation' of various languages.

B. Lindblom: Isn't this a denial of the possibility for a universal framework?

K.N. Stevens: I don't think so. I view these shifts as being fairly small. The high front vowels in various languages may not be phonetically identical, but they are still high front vowels.

K. Pike: It won't work to say it is either 'discrete' or 'continuous'. We need 'particle' or 'wave' descriptions, both of which are observer-related, *and* a 'field' view which describes it in terms of an overall system.

C.J. Bailey and T.V. Gamkrelidze expressed differing views on how much weight to give to typological evidence as opposed to comparative (within-family) evidence when doing reconstructions.

C. Scully: A propos of pre-nasalized stops, I have found in airflow traces that the velum closes very late during the closure portion of post-pausal voiced stops, almost as if some aspects of speech are begun while certain acts of respiration (open velum) are still in play. This may be a good example of a mechanically determined feature of pronunciation that might become generalized and taken up as a linguistic feature.

S. Anderson: I wish to take issue with the assumptions (or by Ohala, an explicit proposal) that claims about phonological structures must be verifiable in terms of substance in some other domain, typically phonetic. At the Phonology session of this congress I sketched a rather different approach to phonology which assumes that there is a systematic domain which is relevant to the nature of language but which isn't directly reducible to other domains. According to this view, the facts that are directly susceptible of phonetic explanations are, in a sense, exactly what is irrelevant to phonology.

F. Longchamp: (To Hombert) You haven't made a clear case for the decreased saliency of the centralized vowels. The vowels that behaved oddly in your study seem to be the one-formant vowels. Of course, subjects can give labels to these vowels but this may have no relevance to natural speech.

H.-M. Jeng: I think child language studies can provide evidence relevant to the questions on the elaboration of segment inventories. In the early speech of my son the consonant system used only the features for ± stop and those for different places of articulation. Later on, features were added to differentiate nasality, aspiration, frication, etc. In the case of vowels, only height features were used at first. Later, front-back and rounding were differentiated. I think these early segment systems represent the universal core upon which further elaborations of the system can be built.

N. Waterson: I question the phonemic basis used in work on universals. There is much evidence that the proper domain of many phonological processes is something more like the word. In sound change the position of the sound in the word and its phonetic context is very important. Children will often produce the correct degree of vowel openness in vowels in a 2-syllable word but not the correct frontness or rounding feature. Thus, when looking for universals we should look for patterns in the domain of the whole syllable or word.

H. Andersen: I don't see how Lindblom's model will accommodate vowel mergers which are very common diachronically. Nor can this problem be solved as recommended by Hombert by assigning the merged vowels to an unnatural transitional state which will eventually revert to a stable natural state. How is one to identify transition as opposed to stable state? The solution, I think, is to recognize that the vowel (as well as the consonant) space is used for more than just diacritic purposes: they also carry information about their consonant environment, about the style of speech used by the speaker as well as his age and social class membership. Thus when the vowels slide around it must be because these subsidiary functions lose their value and are re-interpreted as basic values of the vowel phonemes themselves. This notion is fully in accord with the views expressed here by Profs. Pike and Haudricourt.

L. Jacobson: I can provide some more details on the vowel systems of certain Nilotic languages (alluded to by Ohala) and and at the same time show that they are compatible with Lindblom's model. My own acoustic analysis of the 9 vowel system of Luo shows that many of the non-low vowels show great overlap in an F1 x F2 x F3 space. They can be separated, however, by adding a dimension of voice quality (or pharynx size): breathy voice vs. normal or creaky voice. When this is done, all the vowels are still maximally distant from the other vowels *on the same plane.*

I. Maddieson: It was mentioned (by Lindblom) that high vowels in systems with few vowels tend to be less peripheral. This is a crucial fact and suggests that *maximal* dispersion of entities in an auditory space isn't required. I fihd supporting evidence for this view in the structure of tonal spaces: words borrowed from a 2 level-tone language into a 3 level-tone language reveal that the high tone of the 2-tone language is equal to the mid-tone of the 3-tone

K. Pike: It won't work to say it is either 'discrete' or 'continuous'. We need 'particle' or 'wave' descriptions, both of which are observer-related, *and* a 'field' view which describes it in terms of an overall system.

C.J. Bailey and T.V. Gamkrelidze expressed differing views on how much weight to give to typological evidence as opposed to comparative (within-family) evidence when doing reconstructions.

C. Scully: A propos of pre-nasalized stops, I have found in airflow traces that the velum closes very late during the closure portion of post-pausal voiced stops, almost as if some aspects of speech are begun while certain acts of respiration (open velum) are still in play. This may be a good example of a mechanically determined feature of pronunciation that might become generalized and taken up as a linguistic feature.

S. Anderson: I wish to take issue with the assumptions (or by Ohala, an explicit proposal) that claims about phonological structures must be verifiable in terms of substance in some other domain, typically phonetic. At the Phonology session of this congress I sketched a rather different approach to phonology which assumes that there is a systematic domain which is relevant to the nature of language but which isn't directly reducible to other domains. According to this view, the facts that are directly susceptible of phonetic explanations are, in a sense, exactly what is irrelevant to phonology.

F. Longchamp: (To Hombert) You haven't made a clear case for the decreased saliency of the centralized vowels. The vowels that behaved oddly in your study seem to be the one-formant vowels. Of course, subjects can give labels to these vowels but this may have no relevance to natural speech.

H.-H. Jeng: I think child language studies can provide evidence relevant to the questions on the elaboration of segment inventories. In the early speech of my son the consonant system used only the features for ± stop and those for different places of articulation. Later on, features were added to differentiate nasality, aspiration, frication, etc. In the case of vowels, only height features were used at first. Later, front-back and rounding were differentiated. I think these early segment systems represent the universal core upon which further elaborations of the system can be built.

N. Waterson: I question the phonemic basis used in work on universals. There is much evidence that the proper domain of many phonological processes is something more like the word. In sound change the position of the sound in the word and its phonetic context is very important. Children will often produce the correct degree of vowel openness in vowels in a 2-syllable word but not the correct frontness or rounding feature. Thus, when looking for universals we should look for patterns in the domain of the whole syllable or word.

H. Andersen: I don't see how Lindblom's model will accommodate vowel mergers which are very common diachronically. Nor can this problem be solved as recommended by Hombert by assigning the merged vowels to an unnatural transitional state which will eventually revert to a stable natural state. How is one to identify transition as opposed to stable state? The solution, I think, is to recognize that the vowel (as well as the consonant) space is used for more than just diacritic purposes: they also carry information about their consonant environment, about the style of speech used by the speaker as well as his age and social class membership. Thus when the vowels slide around it must be because these subsidiary functions lose their value and are re-interpreted as basic values of the vowel phonemes themselves. This notion is fully in accord with the views expressed here by Profs. Pike and Haudricourt.

L. Jacobson: I can provide some more details on the vowel systems of certain Nilotic languages (alluded to by Ohala) and and at the same time show that they are compatible with Lindblom's model. My own acoustic analysis of the 9 vowel system of Luo shows that many of the non-low vowels show great overlap in an F1 x F2 x F3 space. They can be separated, however, by adding a dimension of voice quality (or pharynx size): breathy voice vs. normal or creaky voice. When this is done, all the vowels are still maximally distant from the other vowels ***on the same plane***.

I. Maddieson: It was mentioned (by Lindblom) that high vowels in systems with few vowels tend to be less peripheral. This is a crucial fact and suggests that ***maximal*** dispersion of entities in an auditory space isn't required. I find supporting evidence for this view in the structure of tonal spaces: words borrowed from a 2 level-tone language into a 3 level-tone language reveal that the high tone of the 2-tone language is equal to the mid-tone of the 3-tone

language, the implication being that systems with 3 tones use more of the available tone space than do those with 2 tones. We could explain all this as well as the pattern of elaboration of consonant systems by the generalization: additions to these spaces first involve pushing the boundaries of the existing dimensions and then by recruiting additional dimensions for additional contrasts.

L. Lisker: Is the search for universals a viable enterprise if we can't be sure that we are aware of all the features that human languages make use of? New ones are discovered all the time. Also, when making generalizations about segment inventories, we should be clear what we're talking about: the /g/ in English is not the same 'beast' as the /g/'s in Spanish or French, for example. The problem is that the C's and V's we count are invariably the product of the phonologist who uses other than purely phonetic criteria in deciding how to classify sounds.

H. Galton: Considering cases like Ubykh,a Caucasian language with 80 consonants and no more than 2 vowels, and English with about 1/3 as many consonants and many more vowels, I wonder if Prof. Gamkrelidze would accept the tentative universal that is there a kind of balance between a language's consonant and vowel inventories, i.e., that one develops at the expense of the other?

T.V. Gamkrelidze: The number of consonants always exceeds that of vowels since the possibilities for auditory and articulatory contrasts is greater for consonants.

J. Ohala: Regarding the relative merits of a formalist vs. a physicalist research strategy in phonology, the issue raised by Prof. Anderson, I suggest this be decided by examining the 'track record' of the two approaches in providing explanations in phonology.

Reflecting on several of the comments made here, I would suggest we consider the possibility that the single multi-dimensional perceptual space that *both* consonants and vowels range in is not simply defined by the various spectral features (F1, F2, F3), amplitude, periodicity, etc., but rather the first derivative --the rate of change-- of those features. R. Port at Indiana as well as Lindblom have explored this possibility. In this case, the units would no longer be phonemes as such, but rather the transitions between them. These units (more numerous than phonemes) tend to be more invariant, too.

SYMPOSIUM NO. 2: THE PSYCHOLOGICAL REALITY OF PHONOLOGICAL DESPRIPTIONS

(see vol. II, p. 63-128)

Moderator: Victoria A. Fromkin

Panelists: Lyle Campbell, Anne Cutler, Bruce L. Derwing, Wolfgang U. Dressler, Edmund Gussman, Kenneth Hale, Per Linell, and Royal Skousen

Chairperson: Bengt Sigurd

VICTORIA A. FROMKIN's INTRODUCTION

The topic of this symposium is a controversial one. We are hopeful that the debate will lead to new insights and understanding and will help to clarify issues which are important to all sides of the argument. We expect new questions to be raised, questions which we are certain will stimulate the search for answers as to the nature of human language and speech.

Throughout this IXth Congress, the complexities of speech production and perception have been discussed. While we have learned a great deal about these phenomena in the 48 years since the first International Congress of Phonetic Sciences, we still have more questions than answers. The heart of our problem is like that of all scientists, "to explain the complicated visible by some simple invisible." (Perrin, 1914) This is the aim of theory construction, the effort to find a simple, elegant, but "true" (or as close to truth as it is possible to get) accounting of, description of, explanation for the complexities of the phenomena of interest. There is, however, no single approach to how one goes about constructing and validating a theory. That this symposium attests to such differences is revealed in the proceedings (vol. II). We do not even agree as to what constitutes a true theory. The disagreements are, of course, philosophical rather than "scientific". One side of the philosophical debate is set forth by the Nobel prize winning genetecist, François Jacob (1977):

> "... the scientific process does not consist simply in observing, in collecting data, and in deducing from them a theory. One can watch an object for years and never produce any observation of scientific interest. To

> produce a valuable observation one has first to have an idea of what to observe, a preconception of what is possible. Scientific advances often come from uncovering a hitherto unseen aspect of things as a result, not so much of using some new instrument, but rather of looking at objects from a different angle. This look is necessarily guided by a certain idea of what the so called reality might be."

What the reality is constitutes the subject of this symposium. In our case, the reality is a mental or psychological one. We have thus rejected as too confining an earlier definition of linguistics as a classificatory science. (Hocket, 1942) It is no longer enough for a grammar to account for the facts, i.e. the raw data, with the "maximal degree of generalization". The grammar must be a model of the internal grammar constructed by the child; only then will we provide a true description of the language, or a psychologically real grammar.

Even when there is agreement on this aim, different approaches to the job before us are taken. Some linguists and psycholinguists believe that to achieve this goal, it is necessary to test each posited rule in any descriptive grammar to see if it is truly "real". Others suggest that what we are seeking are, rather, constraints on the form of grammars, or a theory of grammar which will answer the question "what is a possible language?" This latter view suggests that with proper constraints any language specific grammar which is permitted by the theory will be psychologically real in that it would be learnable, acquirable by the child when confronted with linguistic data. We all agree that a grammar which is in principle or in fact not "learnable" cannot be psychologically real.

The psychological reality problem did not arise, nor could it have arisen, among linguists such as those who followed Bloomfield in America as they rejected any form of mentalism in linguistics. But even in the early period of the transformational/generative grammar paradigm, the period in which the notion of language as a cognitive system was reintroduced as a legitimate one, there were too few constraints placed on grammars.

I am reminded of the Schachter and Fromkin (1968) phonological analysis of Akan in which final stop consonants /p/, /t/, and /k/ are posited in lexical representation. These

voiceless stops do not surface phonetically in this context. The question that such an analysis poses is whether the Akan child language learner can hypothesize the existence of these final consonants when they never occur in any forms the child hears. Chomsky and Halle (1965) discussed this question a number of years ago.

> "For the linguist or the child learning the language, the set of phonetic representations of utterances is a given empirical fact. His [sic] problem is to assign a lexical representation to each word, and to develop a set of grammatical (in part, phonological) rules which account for the given facts. The performance of this task is limited by the set of constraints on the form of grammars. Without such constraints, the task is obviously impossible; and the narrower such constraints, the more feasible the task becomes."

There are no a priori principles which can tell us what the child is capable of constructing and what she is not. We do not know what the mind is capable of, either the adult mind or the immature mind. In fact, the goal of phonological theory is to provide an answer to the questions concerning the kinds of phonological representations the child can construct, and the rules which can relate these to surface phonetic forms, if indeed there is a difference between these levels. This too is a question for which there is no a priori answer.

The task then of establishing constraints on such a theory such that it will delimit the class of possible grammars to those which are psychologically real, which can be, and which are, acquirable by at least some children, is a task facing us all. If this is the general goal for phonological theory, and let us assume it is, then the question of "psychological reality" is a non-question. We need rather to ask of a theory: "Is it correct?" not "Is it psychologically real?" Or perhaps we should say that the answer to these questions will be identical. In other words, a correct theory of grammars will be a theory of psychologically real grammars.

Unfortunately, even if we agree on this, we find disagreements as to what is meant by psychological reality. I have

therefore asked the participants in this symposium to address this question, to tell us their conception of psychologically real phonological theory.

Closely tied to this basic question are those concerned with the kinds of evidence which can be used to show the reality of a grammar, a lexical entry, an abstract segment, a rule, evidence used to validate or invalidate general theories or particular phonological analyses. In a number of the papers presented in volume II a distinction is made between "external" and "internal" evidence. "External" evidence, as I noted in my summary (p. 63-66), included acquisition data, language disturbance, borrowing, orthography, speech and spelling errors, metrics, casual speech, language games, historical change, perception and production experiments etc. (Cf. Zwicky, 1975) Internal evidence, according to those who make this separation, refers, on the other hand, to facts drawn from the grammar itself, significant generalizations, simplicity factors, distributional criteria, morphemic alternations, etc.

There are linguists, including some of the participants in this symposium, who regard external evidence as more worthy of consideration, as data to be more highly valued than internal evidence. It is not quite clear to me why this should be so. And, in fact, it has been argued that if internal and external evidence are contradictory, internal evidence should prevail. (Cf. below for discussion of Gussman's paper.) External evidence is often performance data, either elicited or observed in actual speech or perception. Speech error data are of this kind. Although I have found, in speech errors, evidence for the independence of features as shown in (1)

(1) Target: Cedars of Lebanon Error: ... Lemadon

where only the value of the feature [nasality] is switched, Klatt (1979) finds "little evidence in the speech error corpus to support independently... movable distinctive features as psychologically real representational units for utterances." While I am not ready to concede to Klatt, let us assume, for the purpose of this argument, that he is correct. Can we conclude from this that a theory of phonology should not represent segments as bundles of features? If we did, we would obscure important

phonological universals in both synchronic and diachronic descriptions; sounds do function in classes, classes which are specified by the features common to their members.

Because the question of internal vs. external evidence has assumed such an important role in discussions on psychological reality, I have asked the symposium participants to present their views on this question.

Each participant has also received one or more questions specific to his or her paper. Let me mention these.

Campbell presents some interesting evidence from Finnish and Kekchi showing the reality of certain posited phonological rules and Morpheme Structure Conditions. He discusses language games played by speakers of these languages. The game data support the rules posited by linguists using internal evidence. Suppose in the language games, these rules were not evidenced. Can one conclude, then, that the P-rules, and MSC's do not exist? That is, what does one do about negative evidence?

This, of course, is not simply a problem that is faced by Campbell, but one faced by all linguists, and, in fact, by all scientists.

Cutler also uses "external" evidence, this time from speech errors, to show that "morphological structure is psychologically real in that English speakers are aware of the relations between words and can form new words from old." She also concludes that "The principles underlying lexical stress assignment are psychologically real in the sense that speakers know the stress pattern of regularly formed new words." This, however, she suggests is in keeping with a "weak" version of psychological reality, which claims simply that speakers can draw on their knowledge of the grammar, as opposed to the "strong" version which would claim that the rules are isomorphic to processes.

It would be interesting to know what kind of evidence would be needed to support the strong version of psychological reality in relation to the posited stress rules of English. What, if anything, does the following error tell us about the psychological reality of the nuclear stress rule?

(Note: for those readers who are not fans of American basketball, Jim West was a famous basketball player with the

Los Angeles Lakers. The meaning of the phrases is paraphrased.)

(1) Target: Jìm Wèst Nìght Gáme. (The game to be played for the special occasion called Jim West Night.)

Error: Jìm Wèst Níght gàme? (the night game played by Jim West.)

Derwing, in his preprinted paper as well as in other of his published works, seems to reject a concept which I hold, i.e. the difference between linguistic knowledge and linguistic behavior. I am therefore interested in how he can find support for psychologically real grammars or rules, given the great variation, including speech errors, false starts, ungrammatical sentences, neologisms, even sounds not ordinarily found in the language that one finds among different speakers of the same language, and even within one speaker on different occasions in both speech production and perception. Is it possible to find exceptionless regularities in behavioral data which permit any generalizations at all? Suppose, for example, one finds five speakers who, to use one of Derwing's examples, relate fable and fabulous, and five who do not. Can we conclude anything? Or should we be constructing individual grammars for each speaker at a single point in time? Or can we conclude instead that, since even one speaker draws certain generalities, the rules which represent them must be psychologically real and permitted by the theory of phonology?

Dressler has distinguished between "naturalness", "productivity" and "psychological reality". How do they relate? Is it possible for a phonological rule to be psychologically real but highly unproductive? And how would such a rule manifest itself. Is there some way that these aspects of language should be delineated in a theory of grammar?

Gussman differs from some of the earlier papers in pointing out that we can not depend on external evidence in our attempts to validate or test phonological hypotheses because it is often the case that different kinds of external evidence are contradictory. It is therefore of interest to know what kinds of constraints he believes should be placed on grammars and how we can find evidence in support of these constraints. Even while he argues that external evidence may be unreliable, he provides such evidence to argue for phonological representations which some linguists would call "abstract". Is this in itself contradictory?

Hale presents a principle which he suggests is needed in a theory of language, the recoverability principle. How is "recoverability" related to psychological reality? Since the principle refers to an evaluation metric for grammars, i.e. a measure by which we can compare the value of grammars, can the metric itself be used to judge whether a grammar is psychologically real? Or, perhaps even more important, how do we judge the psychological reality of any proposed evaluation metric?

Linell gives us a number of interesting definitions. He defines phonology as "language specific phonetics" and rules as "norms". It is thus not immediately clear what the contents of a theory of phonology as distinct from a theory of phonetics would be.

Finally, Skousen has argued that a linguistic description must be directly inducible from the data. At the beginning of this paper I quoted a statement from Jacob which strongly contradicts such a view. The particular paragraph I referred to ends with a further statement: "[Scientific advance] always involves a certain conception about the unknown, that is, about what lies beyond that which one has logical or experimental reasons to believe." Certainly a linguistic description, in the form of a grammar, should be a "scientific advance", an hypothesis, a theory, which goes beyond the collected data. If Jacob is right, why should stronger or different requirements be placed on linguists than are placed on other scientists? And is it possible for us to discover "new truths", to make "new advances" if we are forced to induce all our hypotheses directly from the data?

These are the questions that have been posed for the panelists. We are sure that there are many other questions from the audience which we look forward to hearing.

Whatever our disagreements, we who are the participants of this symposium agree, as I am sure all in the room agree, that

to whatever extent possible we are seeking the "truth", we are seeking a theory of language, and in particular a theory of the sound systems of language, which will bring us a little closer to understanding the beauty as well as complexity of the abilities of the human mind.

References

Chomsky, N. and M. Halle (1965) "Some controversial questions in phonological theory", Journal of Linguistics 1, 97-138.

Hockett, Charles F. (1942) "A system of descriptive phonology", Language 20, 181-205.

Jacob, F. (1977) "Evolution and Tinkering", Science, 196.4295 June 10, 1161-1166.

Klatt, Dennis (1979) "Lexical representations for speech production", paper presented at the International Symposium on the Cognitive Representation of Speech, Edinburgh, July 29 - August 1, 1979.

Perrin, J. (1914) Les Atomes, Paris: Alcan.

Schachter, P. and V. Fromkin (1968) "A Phonology of Akan: Akwapem, Asante and Fante", Working Papers in Phonetics, No. 9, University of California, Los Angeles.

Zwicky, A. (1975) "The strategy of generative phonology", in Phonologia 1972, Dressler and Mareš (eds.) 151-168.

COMMENTS FROM THE PANELISTS

L. Campbell stated his acceptance of the generative phonology goals of descriptive adequacy for particular grammars (which means we should aim at psychologically real grammars) and explanatory adequacy for theories. This requires evidence as to what psychological reality is. Campbell claimed that we cannot find the answer on the basis of internal evidence alone, and one must give greater relative weight to the importance of external evidence. He stated his concept of psychological reality: what is in the head of speakers, i.e. the traditional definition of competence. The more interesting question, he said, is not what psychological reality is, but how do we find out what it is, suggesting that this can only be accomplished by the use of external evidence.

Campbell's answer to the question concerning negative evidence was a simple one: if there is no evidence, there is no evidence. We can conclude nothing. He suggested that a more interesting question concerns counter evidence, which must be used to invalidate theories. He denied the existence of conflicting evidence, despite the reference to such by others. Rather, he suggested that such seeming contradictions are the result of wrong interpretation, theory, or practice.

A. Cutler stated that as she was the lone psychologist on the panel, she would emphasize the "cognitive reality" part of the symposium title by citing some psycholinguistic evidence that prosodic structure is psychologically real. She supported and illustrated her notion of psychological reality by reference to the temporal structure of English, which language is said to exhibit a tendency towards isochrony, in that speakers adjust the duration of unstressed syllables so that stressed syllables occur at roughly equal intervals. She pointed out that there is, however, little evidence that English is physically isochronous; the psychological reality of isochrony is much stronger.

Firstly, English speakers certainly perceive their language as isochronous. In a recent study Donovan and Darwin (1979) presented listeners with sentences in which all stressed syllables began with the same sound, e.g. /t/, and asked them to adjust a sequence of noise bursts to coincide temporally with the /t/ sounds in the sentence. They could hear both sentence and burst sequence as often as they liked, but not together. Donovan and Darwin found that the noise bursts were always adjusted so that the intervals between them were more nearly equal than the intervals between the stressed syllables in the actual sentence--i.e., the listeners heard the sentences as more isochronous than they really were.

Secondly, there is the role of rhythm in syntactic disambiguation. Lehiste (1977) argues that speakers trade on listener expectations by breaking the rhythm of utterances to signify the presence of a syntactic boundary. Durational cues certainly seem to be the most effective at resolving syntactic ambiguities (see, e.g., Streeter, 1978); and recent work by Scott (forthcoming) has demonstrated that boundaries are indicated not merely by

a pause or by phrase-final syllabic lengthening, but crucially by the rhythm--the fact that the foot (inter-stress interval) containing the boundary is lengthened with respect to the other feet in the utterance. Moreover, in a further study of syntactically ambiguous sentences (Cutler & Isard, in press), it was found that speakers tended to lengthen the foot containing the boundary to an integral multiple of the length of the other feet, i.e. "skip a beat" and thus maintain the rhythm.

Finally, there is relevant speech error evidence (Cutler, in press): when an error alters the rhythm of an utterance (a syllable is dropped or added, or stress shifts to a different syllable), it is almost always the case that the error has a more regular rhythm than the intended utterance would have had. In the following examples (syllable omission and stress error), each foot (marked by /) begins with a stressed syllable:

(1) /opering /out of a /front room in /Walthamstow
(Target: /operating /out of a /front room in /Walthamstow)

(2) We /do think in /specific /terms
(Target: We /do think in spe/cific /terms)

The number of unstressed syllables between the stressed syllables is more equal in the errors than in the target utterances. The consistent pattern of such errors supports the notion that isochrony in English is psychologically real: the speakers have adjusted the rhythm of their utterances to what they feel it ought to be.

B. Derwing began his discussion agreeing with Popper (1965) who stresses the importance of the testability of a theory. He then discussed a view which he characterized as that of "autonomous linguistics". According to Derwing, this view holds that there is or may be an idealized natural language system which can be scientifically investigated apart from considerations of the minds and bodies of individual language users. In arguing against such a position, he said that its origins can be traced to a philological notion that a language is an organism complete unto itself and subject to its own unique laws of evolution and change. He referred to a statement of Jespersen that the essence of language is human activity between a speaker and a hearer, and that these two individuals should never be lost sight of if we want to understand the nature of language and of grammar. Jespersen wrote that words and forms were often treated as if they were things or natural objects with an existence of their own. Derwing agreed that such a view is fundamentally false since words and forms exist only by virtue of having been produced by a human organism. For these reasons, Derwing stated he does not embrace the goal of constructing a theory of language, per se, or a theory of possible grammars.

He suggested that modeling the language user is a better goal, since there can be no doubt that speakers learn something when they learn to speak and understand their language, that they know various things as a consequence of this learning, and that they engage in various kinds of internal activity when they put this knowledge to use. The details of this activity and knowledge are amenable to a wide variety of tests. It is thus not the concept of psychological reality which bothers Derwing, but the concept of autonomous linguistics. In fact, he suggested that the question of psychological reality is debated in linguistics only because there are still a large number of linguists who refuse to admit that linguistics is, or at least should be, a branch of psychology.[1]

Derwing stated that only external evidence can provide definitive answers; such evidence is in fact external only from the standpoint of a theory which ignores it. Both kinds of evidence are useful grist for the same mill.

He concluded by saying that it makes no sense to talk of a true theory of natural language since the object of that investigation probably does not exist. The concept of an idealized, monolithic system of language is a notion we can get along very well without. We can, however, subject claims about human linguistic knowledge and abilities to the test of truth. In this enterprise internal evidence is important and suggestive but hardly conclusive.

---

1) In his remarks Derwing did not cite Chomsky (1968) who may have been the first in recent linguistic circles to consider linguistics as "the particular branch of cognitive psychology".

W. Dressler stated that he conceives of psychological reality in the "weak" sense (Cf. Cutler, vol. II, p. 79-85) in that he is trying to account for the competence of linguistic behaviors. His stated approach is to elaborate a deductive theory of natural phonology and a deductive theory of natural morphology, starting from a few basic theoretical concepts. Conflicts concerning naturalness as pertaining to phonology, morphology, the lexicon, etc. would be derived from the theory. Therefore, hypotheses about the psychological reality of these different types of competence would be derived and tested if the intervening variables in each domain of evidence are controlled.

Dressler stated his disagreement with the Chomsky/Halle (1965) statement quoted by Fromkin in which they say the task for the linguist or the child learning the language is similar; the intervening variables for the two are too different for this to be so. Furthermore, he stated that we should not overemphasize child language acquisition at the expense of other kinds of evidence; it is not the privileged domain, and in fact could lead to wrong conclusions. Besides, massive restructuring of the grammar occurs later.

In Dressler's view, external evidence is not extraneous or some sort of supplementary confirmation or disconfirmation, but a central part of the testing procedure. Thus, external evidence can show that an analysis is wrong. He illustrated this with an example from Italian. The masculine article has two forms, il and lo. Phonological and morphological internal evidence suggest overwhelmingly that lo is the basic form. Yet, an Italian asked to give one form in isolation will produce il. Second, the hesitation form, before pause, is il. Finally, change in progress argues for il. These three kinds of external evidence confirm each other and override the internal evidence. The reason is because the techniques for handling internal evidence have mainly been devised for regular phonological and morphological processes and the system of the Italian articles is neither phonologically nor morphologically regular.

E. Gussmann stated that, if phonological descriptions are to be psychologically real, either in the strong or the weak sense, if, that is, they have some kind of correlates in the mind of the user, then the basic question is how we can check or verify the reality of the proposed description. He suggested more caution in evaluating external evidence, pointing to the surprising and, in some cases, contradictory results in direct experiments. Specific examples of this are shown in experiments conducted related to the English regular plural formation rule. In some experiments, subjects responded only 50% in the predicted way, but in others 100% of the forms were those predicted by the regular rule. These experiments say little about whether the English plural rule is productive or psychologically real, but do call for a theory of linguistic behavior which can explain the strange results. What needs to be explained is not only why say, 70% of the answers obtained conformed to the predicted regularity, but, also why 30% failed to do so. In other words, he suggested, one cannot conclude there is no regular rule even when one finds that 30% (or more) responses of subjects in an experimental situation are unpredicted by that rule.

This problem relates to the relative roles of internal and external evidence. Internal evidence, he declared, is primary because it is only in reference to such evidence that external evidence makes any sense.

He went on to discuss the need to reconcile external and internal evidence, pointing to the Dressler proposal for representing the velar nasal in German as deriving from /ng/, and the M. Ohala argument in favor of an abstract schwa in Hindi. It is noteworthy, Gussman claimed, that such cases are usually disregarded by proponents of concrete phonology. Given these abstract analyses, supported internally and externally, one should try to formulate the principles speakers must have access to in formulating such rules and representations. Presumably, he added, one would want these principles to be part of a theory of phonology rather than the phonology of a particular language. It is such principles that we should be seeking.

K. Hale addressed the question of his conception of psychological reality, by stating the question can only be answered when related to the linguist's view of the nature of language itself. In his view, language is a complex human capacity, comprising autonomous, but interacting, systems, each of which has

its own inherent principles of organization. Psychological reality, according to such a view of language, is the goal of linguistic inquiry. It is not given a priori. A logical consequence of this is that it is impossible to ask whether a given linguistic analysis is psychologically real or not, independent of the notion of what is the most highly valued grammar. Thus, the psychologically real, or better still, the most real analysis in a particular instance can only be the one that is best according to some appropriate evaluation metric, functioning internal to the particular framework in which a particular analysis is cast and resulting in some natural way from that framework. He added that, in his candid and probably unpopular view, the traditional generative grammarian's notion of a simplicity metric is on the right track. The problem is to have the right metric, no simple matter.

In discussing the question of internal vs. external evidence, he said he finds it difficult to make the distinction, preferring to distinguish between good and bad evidence. When a field linguist is faced with two or more possible analyses of some data, (s)he needs to look at any kind of evidence to decide. In the case of the Maori passive which he discussed in his paper (vol. II, p. 108-113), the analysis he arrived at after looking at ten different kinds of evidence was the unexpected one, setting up a conjugation system among verbs rather than presenting a purely phonological analysis. Yet the phonological rule analysis would probably be the one required of any student who wanted to pass a phonology course. Hale argued that strictly linguistic reasons favor the morphological analysis, referring to Jonathan Kaye's "recoverability principle". This principle also appears to operate in Papago, to select an analysis which could be considered to be just the opposite from that in Maori, although the surface phenomena are identical. This principle may then be a subcase of a more general simplicity metric, affirming the importance of such linguistic principles. He concluded by stating that the psychologically most real analysis will be that most highly valued by a valid simplicity metric.

P. Linell argued for a behavioral performance perspective on language, stating that a language should be viewed as a system

of grammatical and phonological phonetic conditions placed on the stream of meaningful and phonetic communicative behavior. He thus would assign a role to phonological form both as related to plans for the pronunciation of the expressions in question and as related to perceptual schema. Phonological entities are phonetic entities, i.e. phonetic behavioral articulary plans, intentions, perceptual schemas etc. There are phonological aspects of morphological formation patterns which he said also belong to other components of the grammar, but these, too, concern surface phonetic entities.

Linell suggested that whether one considers psychological reality a non-issue depends on one's theoretical preference. If a language is seen exclusively as a set of abstract sound-meaning correspondences, isolated from behavior and communication, it probably is. Thus, he maintained, autonomous linguistics aims at capturing all detectable generalizations at all levels, and this is a legitimate concern. But if one is interested in psychological reality, Linell proposed that it is necessary to look at production and perception behavior, language learning, and language storage. A language user does not need all the linguists' generalizations and it is thus doubtful that these are psychologically valid. It is more likely, he claimed, that there is great redundancy in the grammar leading to processing short cuts, heuristic routines, parallel strategies etc.

In arguing against formal conditions on rules, or principles, he stated that too often such discussions are pointless since when, for example, we raise the question of recoverability, why should morphophonemic forms be recovered at all, by whom are they supposedly recovered, and for what purpose.

The problem cannot be solved by experimentation, he added, unless we know how to interpret the hypotheses we are testing. If, for example, we find speakers make the vowel substitutions predicted by the vowel shift rule in SPE, we should not conclude that the way the rule is formulated is correct. (Chomsky & Halle, 1968) Or if speakers relate fable and fabulous it is a non-sequitur to conclude that there is one morpheme form underlying both words. This is the generative way of describing the relationship, but there are other possibilities.

Linell concluded with the suggestion that it may be artificial to separate out psychological reality from social and biological reality. What we want is a true synchronic theory of the linguistic practice of language users.

R. Skousen suggested that the psychologically real descriptions which we seek may not be composed of rules such as the kind that have been postulated, or any rules at all. Although linguists may characterize behavior in terms of rules, it is not certain that linguistic behavior itself is rule-governed.

He illustrated his point of view by a discussion of "probabilistic" rules. He considered a hypothetical language in which the verbal past tense is realized by one of two forms, in what has been called in the past free variation. But, suppose in observational studies it is found that a given speaker produces one of these forms two thirds of the time, and the other, one third of the time. He provided reasons why one should not posit a rule which specifies the probability of occurrence of either form in that speaker's grammar. A linguist can construct such a rule, but this does not mean that a speaker can or does construct a rule of this form.

He followed up this example with a discussion on apparent regular rules with exceptions and questioned whether in many of these cases we should conclude that the speaker utilizes a rule rather than looking for specific forms and then using these forms analogically to produce new and novel forms.

DISCUSSION

A discussion ensued, participated in by the panelists and by the following speakers from the audience: C.J. Bailey, R.P. Botha, J. Bybee Hooper, R. Coates, T. Gamkrelidze, W. Labov, A. Liberman, L. Menn, J. Ohala, and J. Ringen. There will be no attempt to cover all the interesting points presented.

A number of the discussants continued on the topic of internal vs. external evidence. Ohala posited that this is a false dichotomy, a point made earlier by Hale, since evidence is evidence. He suggested, however, that there is a continuum in the quality of evidence, since some evidence may be less ambiguous and more capable of refinement than other evidence. He stated that "internal evidence" is highly ambiguous as to what it reveals about psychological entities; evidence from speech errors is of slightly higher causality, and evidence from experiments the least ambiguous and the most capable of refinement because of experimental controls.

On the same question, Bybee Hooper referred to the external evidence used to support the velar nasal as deriving from /ng/ and said that there are other interpretations which can be made, thus warning against making unwarranted assumptions about linguistic structure from such evidence. Both Gussman and Campbell agreed that unwarranted assumptions shouldn't be made about anything.

Hale pointed to the possibility that there may be opposing analyses for which no external evidence is available, and suggested that it is highly possible that a child confronted with a language has a problem similar to that of the field linguist who has only the language data. He suggested that we therefore need some internal principles which permit both the linguist and the child to come up with an analysis. He pointed to problems in interpreting external evidence like that of language games. He has found that in Australia, where secret languages are elaborate and a key intellectual activity among the aboriginal people, some are very good at these games and others very bad. Thus one gets variable data.

Labov followed the lead of Linell's suggestion that one must consider other forms of reality such as social reality, and, in fact, argued that this may have greater importance than psychological reality. He pointed to evidence from child language acquisition showing that children use different strategies before their grammars converge, and he said such differences probably persist in the more irregular portions of the language for some time. In his study of Philadelphian English, he has found that some Philadelphians use a complex rule to derive two phonetic vowels, whereas for others, it appears, two underlying forms exist. Much of the evidence we seek refers to the social reality of the system rather than the processing of individuals.

Bailey also considered the importance of language change, going so far as to say a dynamic approach must be used rather

than a static one in looking at language.

Campbell also added to the discussion on social factors by pointing to the fact that they can complicate phonological descriptions. He has found that in some societies the avoidance of "dirty words" causes phonological complications. Dressler noted that considerations of social reality and the social and communicative function of language was key to a concern for universals in phonology. In discussing variation across individuals Derwing noted that sociological reality was nothing more than a sum of the psychological reality of many individuals. If, he said, we are studying language users, we do not expect them to be the same. Linell suggested that rules should be construed as socially acquired and socially shared, which, he added, is the traditional notion of a rule as a norm for behavior.

Ringen and Botha both discussed the role of the philosophy of science in theory construction and validation. Botha stated there is no such thing as the problem of psychological reality of phonological descriptions. There may be a problem, and this depends first, on the aims of the theory, and second, on the philosophical approach of the linguistic scientist. The notions of "truth","reality", and "evidence" are theory bound. Ringen also noted the relevance of philosophical questions. He also affirmed the importance of theories of performance in deciding whether evidence is internal or external.

Cutler also argued for the need for a theory of performance but, as a psychologist, pointed to the difficulties in attempting to set up psychological experiments which would get at the strong version of psychological reality. Coates also stressed the importance of working with psychologists in our attempts to establish the kinds of association between linguistic units which exist. The notion of units was discussed by Liberman, who stated that the basic task for phonology is to segment the non-discrete speech signal into the correct discrete segments.

Gamkrelidze noted that the goal of constructing a theory which would provide for psychologically real grammars was not one which arose with the transformational linguists, who, instead, he believes placed their emphasis on cybernetic considerations. He pointed to the difficulties, however, of trying to determine what is in the mind of speakers, from their utterances, which parallels the difficulty of trying to determine the inner mechanisms of a clock from watching the hands move. Many models can be constructed which give the same output but only one model is the correct one. This point was similar to one made by Skousen in discussing the need for real world interpretations of formal linguistic constructs, providing an interesting analogy with a formal system of Euclidian geometry which can only have "reality" when the formal primitives are given substantive interpretations.

Menn was concerned with the fact that linguists, or some linguists, seem to ignore the variety of things which can legitimately be considered knowledge and the necessity of distinguishing among them. SPE ignores the degree of rule productivity, she noted, and most experimental linguists ignore the difference between active and passive knowledge and the difference between explicit metalinguistic knowledge ("I can tell you that word A contains morpheme B") and implicit knowledge ("I guess that word A is more likely to mean something about rocks than sugar.") We need to set up sufficiently subtle experiments to be able to differentiate between these phenomena, she said.

To conclude the symposium, the moderator, Fromkin, presented some of her own thoughts. She agreed that it is not possible to proceed without any biases or a specific philosophy of science. One would hope, however, that despite different philosophies, linguists will provide increasing information which will reveal something about the phonological systems of the languages of the world.

She referred to some of the arguments concerning "autonomous linguistics" and expressed confusion as to what that phrase really does mean, or why some people consider it negatively. No one can deny that language is used in society, that language is a product of evolution, that there are brain mechanisms underlying language, that language is used by speakers in producing utterances and in comprehending speech, that it is used for humor, for making love, for expressing hate, for selling soap, but, she asked, why is it not legitimate to attempt to study the language systems which underlie all these uses, to investigate language

per se. The history of science shows the isolation of different facets of reality in order to better understand them. Do we need to study the persuasive and disgraceful use of ambiguities by advertising agencies before concluding that for some speakers of English writer and rider are homophonous even though write and ride are not? And that the homophony arises from an "alveolar flap rule"? Whether or not one believes in the reality of rules, in describing the sound patterns of English, we certainly must reveal this "fact".

This does not mean, she added, that we can ignore the bridges between one part of the complex phenomena and another. But it certainly is legitimate to say that human language exists and we should try to understand it. The question then arises as to whether language is a cognitive system which can be viewed apart from the behaviors of those who have acquired it. Those who hold this opinion point to various kinds of evidence to support it. For example, many if not all of us produce utterances which we, in hearing a tape of our own speech, will regard as "improper" or ungrammatical. This judgment must come from some stored knowledge. Clearly we can and do say, produce, and understand the meaning of utterances that we also declare to be ungrammatical sentences. Thus utterance is not equal to the theoretical construct, sentence.

Fromkin continued her discussion on "autonomous linguistics" saying that the pursuit of language per se may be a worthy one. This does not imply that linguistics is not a subset of psychology. Derwing's dichotomy does not necessarily hold, if we view language as a system of knowledge that is a mental reality. There are of course many subsets of psychology. One can pursue research in the field of vision without conducting research on auditory perception. Furthermore, psychology is concerned with behavior but not exclusively so. There are as many differences of opinion among psychologists as there are among linguists, many stemming from differing philosophical views. Fromkin stated that she could probably point to as many psychologists who agree with her view of the aims and proper subject matter of linguistics as can Derwing in support of his views.

However, she wished to emphasize that this does not mean that the construction of performance models is not a worthy one for linguists. Her own research has been primarily concerned with performance, but she added that this research has been guided by the insights provided by linguists working on language structure, rules, and representations.

Failure to distinguish between linguistic behavior and knowledge would create problems for those analyzing speech errors. Similarly, the study of aphasia shows that in many cases the linguistic deficits are performance deficits, while the stored grammar is intact. Otherwise one could not explain why an aphasic patient is capable of production, retrieval, and perception on one day, and incapable of one or the other aspect of performance on another occasion. Manfred Bierwisch pointed to this discrepency many years ago when he posited that most aphasia symptoms can only be explained as performance breakdown.

Fromkin concluded with a quote from Poincaré (as cited in Chandrasekhar, 1979):

> "The scientist does not study nature (only) because it is useful to do so. He studies it because he takes pleasure in it because it is beautiful. If nature were not beautiful it would not be worth knowing and life would not be worth living."

She ended by saying that we who are interested in human language know how meaningful this quote is, since human language, like all of nature, is beautiful, and the study of it is therefore worth doing.

References

Chandrasekhar, S. (1979): "Beauty and the quest for beauty in science", Physics Today, July 1979, 25-30.

Chomsky, N. (1968): Language and mind, New York: Harcourt Brace Jovanovich, Inc.

Chomsky, N. and M. Halle (1965): "Some controversial questions in phonological theory", Journal of Linguistics 1, 97-138.

Chomsky, N. and M. Halle (1968): The sound pattern of English, New York: Harper & Row.

Cutler, A. (in press): "Syllable omission errors and isochrony", in Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler, H.W. Dechert (ed.), The Hague: Mouton.

Cutler, A. and S.D. Isard (in press): "The production of prosody", in Language production, B. Butterworth (ed.), London: Academic Press.

Donovan, A. and C.J. Darwin (1979): "The perceived rhythm of speech", Proc.Phon.9, vol. II, 268-274, Copenhagen: Institute of Phonetics.

Lehiste, I. (1977): "Isochrony reconsidered", JPh 5, 253-263.

Popper, K.R. (1965): Conjectures and refutations, New York: Basic Books.

Scott, D. (forthcoming): Perception of phrase boundaries, Ph.D. Thesis, University of Sussex.

Streeter, L.A. (1978): "Acoustic determinants of phrase boundary perception", JASA 64, 1582-1592.

SYMPOSIUM NO. 4: SOCIAL FACTORS IN SOUND CHANGE

(see vol. II, p. 185-237)

Moderator: Einar Haugen

Panelists: Henrik Birnbaum, Ivan Fónagy, William Labov, Jørn Lund, Bertil Malmberg, and Fred C.C. Peng

Chairperson: Martin Kloster-Jensen

EINAR HAUGEN's INTRODUCTION

1. The Contributors and their Papers. Each of the invited speakers in this symposium has done research and thought deeply about the topic of linguistic change. They range from newcomers like Lars Brink and Jørn Lund to elder statesmen like Bertil Malmberg. It is one of the prime purposes of such congresses as this to bring together representatives of different views, different ages, and different countries, so that their ideas may be discussed face to face. Unfortunately, each contributor is limited by the format of the occasion to a short presentation in print of the main results of his research and an even shorter presentation by word of mouth. My function as moderator has been the pleasant one of summing them up and showing how together they constitute an advance toward our understanding of the central problem that is the topic of this symposium. One difficulty is that the authors deal with many situations that I do not know firsthand, and that they take up different aspects of the problem itself. In some cases I have had to go back to other work by the same and other authors to clarify the problem in my own mind.

2. Theorizers and Empiricists. The contributors fall into two categories, which I shall call "theorizers" and "empiricists". The "theorizers" are those who base their discussion largely on informal observation from which they make more or less intuitive generalizations. This is not a pejorative description, for in this field I count also myself. I would count among them Birnbaum, Fónagy, and Malmberg. The others are "empiricists" because they present actual field work, much of which has been statistically treated, so that their conclusions give the refreshing impression that we may be able to treat an old problem in a new way, namely by direct observation. I find this approach most exciting, since it builds on forms of data gathering that have become possible once we had tape recorders, computers, and spectrographs. Phonetic

change used to be considered as something we could observe only over centuries. We are now told that we can catch it on the wing. Instead of observing its results only, we can now see it going on. This development appears especially in the papers of Brink and Lund, Labov, and Peng. It has made possible an empirical sociolinguistics, of which earlier investigators could only dream.

3. The topics. I shall first present a very brief statement of the contents of each paper, beginning with the theorists. Birnbaum is largely concerned with criticizing a linguistic model of decoding advanced by Henning Andersen under the name of "abductive". He does not believe that it can account for the rise of innovations in a homogeneous speech community, a construct which in any case he rejects. Fónagy is here concerned primarily with intonation and its historical development. He rejects all notions that it is a "universal" or that it is a fixed, non-arbitrary and motivated phenomenon. Malmberg sees a "state of language" as "a harmonious achronic system or rather complex of systems" within which the speaker may choose according to situation. His chief example, which he has previously studied in detail, is the Parisian vowel system, or rather its "maximum" and "minimum" systems. He regards the rise of "minimum" systems as the result of a "simplification" that is typical of persons living on the social and spatial periphery of a society. Brink/Lund (as I shall call them jointly) have gathered a vast amount of data on the phonetics of Copenhagen speakers born between 1840 and 1955, fully presented in their massive two-volume Dansk Rigsmål (Copenhagen, 1975), unfortunately available only in Danish. Basing themselves primarily on phonograph recordings going as far back as 1913 as well as whatever printed materials are available, they have identified up to sixty regular phonetic changes. They have divided their speakers into two social groups, speakers of "high" and "low" Copenhagen. Labov's work has dealt with a variety of American groups, beginning in the island of Martha's Vineyard in Massachusetts, continuing on New York's lower East Side, and currently in Philadelphia. He has concentrated on Black youth, but has worked with all colors and social classes. Finally, Peng bases himself on extensive data gathering in Tsuruoka, Japan, by his colleague Nomoto. This was a sample first drawn in 1950 and then reexamined in 1971. The novelty in his theory is that one generation is sufficient to

identify the process of sound change. Labov's period is in some sense even shorter, since he studies different age groups synchronically and assumes that young people will carry their innovations on into adulthood. We are fortunate in having a wide variety of data bases, from three continents, as well as considerable variety in theoretical approaches.

4. Stability and Change. Except in immigrant communities, every community studied so far has enough stability of language so that each generation can communicate with every other. At the same time language is known to be changing at a rate such that after some unspecified number of generations it will become unintelligible to its ancestors. These basic facts determine the possibility of two complementary views: that language is stable and can form the object of synchronic study, and that language is constantly changing so that it can form the object of diachronic study. In their extreme form both views become unrealistic, e.g. in assuming complete homogeneity or complete fluidity. Members of the Prague School (e.g. Havránek, see Garvin 1959) described "elastic stability" as desirable in a standard language, but in fact they were only defining the nature of all language, "standard" or not. Labov has invented the latest synonym for this term in his "orderly heterogeneity", which is as much a construct as Chomsky's "ideal homogeneity" to which he opposes it. Both agree that language is "structured", i.e. amenable to description by rules. Chomsky's are categorial, Labov's variable, but there is structure in both. The step from categorial to variable rules is a great step forward in descriptive linguistics, but it was foreseen in historical linguistics, and especially in dialect geography.

Here it is useful to emphasize the concept of "choice" as used in Malmberg's paper. Variable or conflicting rules mean that individuals have the freedom to change language within wider or narrower limits of acceptability. But none of these rules are very helpful so far in predicting the future. Any attempt to predict sound change has to face the problem of showing why people make decisions as they do. But this involves going back into their individual and collective psyches to study their unconscious motivations, an infinite regression that leads us far outside the realm of most linguists' competence, though some have loved to

speculate about it. A careful study of the tiny rule changes in Copenhagen speech pinpointed by Brink/Lund suggests that at any given moment in time there is an enormous amount of unstructured heterogeneity, of vacillation and uncertainty. This may either continue, or be resolved by a later generation, and it may lead either to innovation or to regression.

5. The Problem of Actuation. It is hardly surprising that living language abounds in heterogeneity. It is more surprising that there is no more of it than there is. The basic reason for heterogeneity has been evident ever since men stopped believing in such myths as the Tower of Babel. Recent linguists have rediscovered the fact that language is innate and universal, but the most universal fact about languages in the plural and concrete is that every one of them has to be learned anew by every human being born on this planet. He or she is born to human parents and in a human society, surrounded by the speech output around it. That output becomes the input to the child's own processing of the language for reception and eventually production. The study of the child's language learning (which for some arcane reason has come to be known as "acquisition" -- perhaps it is part of our acquisitive civilization) has become an important field of research. We may look to its results for new light on the extent to which the fully formed child's language differs from that of its environment. We do know that eventually all non-defective children learn to communicate in whatever language variety is spoken around them, in spite of the inevitable differences among individuals in talent, appearance, industry, and success. But human beings are not robots and no given language is imprinted by instinct. Try as they will, people will deviate. Call their deviation a "speech error" or a "creative innovation", as you will; it is the germ of a language change.

6. The Mechanism of Diffusion. Given the fact that more or less random innovations occur, we need to pinpoint the process by which they are spread to other speakers. If they fail to spread, they remain speech errors; if they do spread, they become linguistic changes. On this point our symposium speakers show a clear difference. Brink and Lund appear to believe that the innovations are made in childhood and are then retained for life, unless of course the speaker moves into a new linguistic environment. Their basis for this claim is the recordings they have studied of the same speakers at various periods of their lives. It must be noted, however, that age 15 was the lowest they studied, which is already after the onset of puberty. Many studies have shown, whatever the cause of it may be, that puberty is a period when language tends to fix itself into an adult pattern that most people find difficult to change. Birnbaum emphasizes the importance of the teens as "the age when growing-up speakers, by imitating their elders, attain the same or nearly same pronunciation as their models." He regards such changes as frequently deliberate, and due to fashion within the generation. At the same time he rejects the simple transfer of one generation to another, since there is a "continuous pattern-setting effect of parents on children, teachers on students, leaders on followers, older on younger playmates and fellow workers, more prestigious on less prestigious..."

Against this view Peng entirely rejects the idea that change takes place across generations. He specifically denies Johnson's (1976) view of an accelerating change over three generations. He has found that Nomoto's speakers showed many changes over a period of 21 years. He suggests that while the rate of change may go down as age goes up and reaches a low point around age 35, it never completely stops. He questions Labov's use of "apparent" time studied in synchronically present generations and advocates the use of "real time". Presumably Labov would agree that this is desirable when the investigator lived long enough, or when his informants do, for he (Labov) refers to Hermann's restudy of Gauchat's famous village of Charmey in Switzerland. Peng suggests as an alternative the use of dialect geographical material, with its mapping of horizontal linguistic change. This, too, is a case of apparent time, however, since the dialects exist synchronically, and we can deduce just how or even approximately when the change took place only by the use of comparative-reconstructive methods.

7. Class Correlations. Our speakers also show certain differences of opinion concerning the role played by social and other classes in the actuation of change. Labov has found that in American cities the upper working or lower middle class, that is, the centrally located classes, lead in linguistic change. The speakers who are most advanced are the ones with the highest aspirations for advancement, who also have the largest number of

local contacts outside the community. Malmberg has fixed his view on the central norm of Parisian French and regards simplification as a major factor, which he then attributes to the lower classes and the provincials, who live on the periphery. In Brink/Lund's detailed account of their three-score changes in Copenhagen, however, the role of social class is rather different. To begin with, they deny that there were what we would call class differences prior to 1750. Before that time the speech of Copenhagen was a local dialect like any other, different from its neighbors, but having much in common with them. In the 18th and 19th centuries a class differentiation took place which reduced contact between different strata of society. A distinct lower-class speech developed, which in general was ahead of upper-class speech. Only since 1900, when everyone is sending their children to publicly supported common schools, are the differences leveling out, or in the view of the élite, the language is being "vulgarized". Unfortunately, it is difficult to compare Brink/Lund's results directly with Labov's, since they operate with only two classes as against Labov's more refined indices of class membership.

On one point everyone seems to be agreed: that women everywhere are more "refined" than men of the same age or class, i.e. have more features classified as "high". Brink/Lund are not willing to grant the existence of a separate "sexolect", but suggest that women are more sensitive (perhaps rather "sensitized") to social status. Fónagy finds that in Hungarian a final rising intonation has lost its marked value as an indicator of "expressiveness". The reason is that it has now become normal among women and young people.

8. Conclusions. Two of our speakers emphasize that it is not language that changes, but people who change language. Peng writes, "People change, and sound change is simply a manifestation (or symptom) of human change." Malmberg reiterates from his Bucharest paper (1969) that "language does not change; man changes languages." These statements are true, but tautological, unless we are speaking of the adoption of new words or the learning of new languages. Phonology tends to fall below the threshold of consciousness for most speakers, and they are rarely aware of making changes in their own speech. It is only with the greatest caution that we can identify any external social reason for such

unconscious change. Nothing in climate, occupation, physiology, character, or history can be causally connected with such large-scale linguistic changes as the Germanic consonant shift, or Umlaut, or the English vowel shift, or even with the decay of inflections in most Romance and Germanic languages.

Brink/Lund even deny that the Copenhagen forms have spread because of the prestige of the capital city. But their claim that they spread "purely by contagion" makes one wonder why they did not spread the other way, during a period when the city was invaded by great numbers of rural immigrants. They believe that new pronunciations spread by virtue of an "inherent plus value", vaguely defined as their being "easier to articulate", and conclude that "sound change is essentially a non-social phenomenon." William Labov, who has done more to correlate social and linguistic variation than anyone else, is equally pessimistic: Bloomfield's assertion of 55 years ago that "the causes of sound change are unknown" is still true.

In spite of the weight of first-hand research and authority which these writers bring to the topic, I cannot let this conclusion stand as the final word of the symposium. I am convinced that the causes are known, but that what is really meant is that the results are unpredictable. Let me briefly sum up my own unsupported and intuitive view of sound change (though it is not unlike that held by Hugo Schuchardt and Otto Jespersen). Sound change is in principle no different from any other change going on in the lives of animate beings everywhere around us. To say that we do not know the causes of change is like saying that we do not know the causes of human fashions, e.g. the length of women's skirts or the shape of men's headgear. We do know that one main cause of human language change is that language is not genetic, but learned, and that no two human beings ever learn anything exactly alike. I do not believe that the parts of any language hang together in Meillet's sense of "tout se tient". If they did, there would neither be sound change nor the development of dialects. I believe instead in what I may call the "amoeba" theory of language, that any aggregation of items we call a "language" or "dialect" is as arbitrary as the movements and splittings of the amoeba. The most important rules of language are simple collocations. Phonetic changes can only have been "actuated" by

individual learners and users, whether as children or adults, who committed errors in hearing or reproduction that were not corrected by themselves or others. Phoneticians can tell us a great deal about the physical and acoustic parameters that favor such errors, but they cannot predict which of them will occur.

To become part of the speech of others, these innovations have to be acceptable to other members of the community. This is the process of diffusion, which has to be both lexical and social. Lexically, the change has to spread from the one item in which it started to other items that in some way are felt to be similar to the first. The neogrammarians' or any other linguistic formulation of such changes or "rules" as they are now called is an ex-post-facto summary of change, not a description of the change itself. As dialect geography clearly shows, a change may stop at any point in its diffusion, before it has spread to the entire lexicon or the entire community. It may even change its domain, be reordered or reorganized, apply to different parts of the system, be lexicalized or grammaticized. "Simplification", which is often resorted to as an explanation, is no real answer, for neighboring dialects fail to simplify in the same way. According to Chen and Wang (1975:267), the final nasal consonant /m/ has been lost in Mandarin, but in Cantonese it is still there. Who could have predicted that? It is vocalized in French, but in English we still have it. A tendency, yes; a universal, no. Besides, in spite of all simplification, every language known seems to be of about equal difficulty, learned at much the same age by children who are exposed to it.

There are too many factors present in every human situation for us to be able to foresee all its possibilities. No sooner has one rule operated for a time than another takes over and messes it up. Such is life, and language is no different.

References

Chen, M.Y. and W.S-Y. Wang (1975): "Sound change: actuation and implementation", Language 51, 255-282.

Fónagy, I. (1956): "Über den Verlauf des Lautwandels", Acta Linguistica Academiae Scientiarum Hungaricae (Budapest) 6, 173-278.

Garvin, P.L. (1959): "The standard language problem -- concepts and methods", Anthropological Linguistics 1, 3, 28-31.

Jespersen, O. (1933): Linguistica: Selected papers in English, French and German, Copenhagen.

Johnson, L. (1976): "A rate of change index for language", Language in Society 5, 165-172.

Labov, W. (1972): Sociolinguistic patterns, Philadelphia.

Malmberg, B. (1969): "Synchronie et diachronie", Actes du X^e Congrès International des Linguistes, vol. I (Bucarest), 13-25.

Schuchardt, H. (1928): Schuchardt-Brevier, Halle.

Sommerfelt, A. (1968): "Phonetics and sociology", in Manual of phonetics, B. Malmberg (ed.), 488-501, The Hague.

COMMENTS FROM THE PANELISTS

Birnbaum did not intend his paper as a major critique of Henning Andersen's abductive model of phonological innovation, for which he has great admiration. He only wished to indicate that it could be improved on some minor points, e.g. the problem of generational sequence. He was concerned with any trend toward excessive schematism. As for being classified as a "theorizer", he wanted to make it clear that he believed in a happy combination of data gathering and theorizing. He agreed that early childhood was the most important period for establishing speech habits, but that puberty also led to readjustments.

Fónagy was stimulated to study French accent after being rebuked for having an 18th century pronunciation on his arrival in France thirty years ago: he made it a habit to place every stress on the last syllable! He has found that French stress is elusive: its placing is a probabilistic function of many variables, including syntax, genre, etc. Today radio and television speakers are increasingly stressing enclitics, which are not stressed in conversational speech.

Labov described his paper as the first report on his Philadelphia study, his largest project so far, using more advanced techniques than his earlier studies. He has adopted the strategy of searching for innovators: where are they in the social spectrum, by sex, class, position etc. How is sound change related to the network of communications and to new ethnic groups that enter society? Can we throw light on change by looking at the people who are doing it? He does not think that the individual is a significant unit: we are dealing with the social pressures which form an individual into a social being as he grows up and assumes

a variety of roles in the social structure. His main motivation in coming to the meeting was to make contact with acoustic phoneticians and the theoreticians who have developed the models we use: Fant, Fujimura, etc. "Ever since 1968 we've made the point that the tools of acoustic phonetics are useful for examining problems of language structure and language change." These tools will require increasing understanding of the mathematical models at the base. A report on the Philadelphia study should be available in three or four months.

Lund, on behalf of himself and Brink, spoke about their findings in the study of Copenhagen pronunciation. They found that "the sound pattern of the single individual will not change significantly after the teenage years unless the linguistic environment is changed rather profoundly." In the book they had taken the position that sound change takes place across generation boundaries, but they did not deny Peng's contention that sound changes in progress can be studied within one generation. But in this case there is often situational variation, with old forms in more formal speech, new forms in more casual speech. Here Malmberg's distinction of maximum and minimum may be applicable, though they found the term "minimum system" problematic. In casual speech there are not only the typical reductions and assimilations, but also subconscious new sound qualities that do not necessarily lead to simplification. Nor can they see anything here in common with aphasic speech or the reduced inventory of phonemes often characteristic of foreigners. They agree with Fónagy that changes in prosody "must be accounted for in the description of linguistic evolution." They question Labov's finding that the most advanced speakers "are those with the highest status in their local community, having found that new pronunciations have low prestige and are often considered vulgar, if noticed at all." They agree with Haugen that most changes are unconscious and that their investigation is difficult to compare with Labov's, since they started from the phonetic variation, and only secondarily examined the social correlation. "No Danish pronunciations are characteristic of the middle classes."

Malmberg noted that his paper "starts from my distinction between a language as a closed, hierarchical system of mutually dependent units, a structure sui generis, in our case the phonemic

system, where any change in the number and/or the relations of these units implies the creation of a new language richer or poorer or differently structured." Further, "a state of language..." (a Saussurean term) "is a sociolinguistic concept which for its full definition needs extra-linguistic parameters." "Every system or subsystem ... can function as one of the layers within a state of language." "The degree of mastery and retention of the complexity...is a question of the strength of the social norms which determine the speakers' behavior. The terms 'maximum' and 'minimum' systems must be understood as abstractions." By "simplification" he referred to phenomena occurring in the social and geographical periphery of normative centers and areas in contact with other systems on the linguistic border, including the diffusion of languages to new areas through colonization. He did not have in mind peripheric local dialects, which can be very conservative. "My principal point is the existence of layers of varying complexity and of norms of varying strength and the (socially determined) choice between different possibilities." "My intentionally provocative formulation at the Bucharest Congress in 1967 was made to stress the importance of the choice factor and that of social evaluation in phonetic/phonemic change."

Peng called attention to the two basic assumptions in his paper: (1) That language change is a change in behavior. Only by studying changes in language behavior can we discover changes in the code. Once this step is taken, one can observe changes within a single generation, without waiting for two or more demographic generations. (2) A random sample is more representative of human behavior than one that is previously stratified for class. In his work in Tsuruoka the same questionnaire was administered to 137 informants chosen at random and interviewed 21 years apart. In this way it was possible to make use of real rather than apparent time. In plotting the changes over time, one gets a straight line, showing that all age groups were affected.

Labov agreed that people tend to preserve their vernacular and gave the example of a mother and a daughter who differed widely in the pronunciation of the /aw/ diphthong. But he granted that people change their norms and only now realized that Peng had been studying the formal responses to norms and not the vernacular. He himself was looking for un-

reflecting speech, "the most systematic motor-controlled speech." No one has studied syntactic change, which may indeed be individual (cf. study by René Agneau of the progressive in 19th century English, showing that e.g. George Eliot made increasing use of the progressive in the course of a half century.) He expressed admiration for Peng's use of real time, but in his own work he preferred to begin with people in the context of their local community. He agreed with Lund that whenever changes rise to the level of consciousness, speakers tend to reject them.

Birnbaum commented on the moderator's summary. He gave an example of women's speech as different from men's: women tend to use an implosive /h/ in a word like jaha. He agreed that prediction is dangerous, and gave an example from Polish, the replacement of nasality in final vowels by diphthongization. Also that we can ascertain the causes of change, but that we cannot always explain them. He found the summary to be an important paper, by virtue of the moderator's including views of his own, perhaps unduly pessimistic.

Haugen as moderator responded modestly that he found the non-systematic parts of language more interesting than the systematic ones, whose existence he had never denied. He found that only by assuming an arbitrary disjunction between the parts of a system could one explain that they could change independently. One example is the well-known fact that an adult learner can speak a language fluently and with virtually perfect syntax and lexicon without ever mastering the phonetic system.

Peng noted that he had speculated on the causes of change and found many factors and mechanisms. He did not feel that the generation boundaries were primary, but the fact that speakers pass on a different language from the one they themselves learned. Diffusion of the code and diffusion of the people who accept it are two concurrent dimensions of diffusion. He challenged Lund to explain how he arrived at his conclusion of non-change on the part of individuals.

Fónagy mentioned retrospective studies of linguistic change in the 16th-18th centuries. They show that there are enormous differences between sound change and sound change. Some changes are dependent on sex (one reason given for a difference in women's speech at that time was that it was not good form for them to

open their mouths too wide), others are not. Some changes are socially dependent, some are word class dependent, others are not.

Lund replied that they had made spot checks of the same person recorded in the same speech situation many years later.

DISCUSSION

Simone Elbaz (Paris): "Mon intervention n'est en rien polémiste. C'est une mise au point. J'ai le plus grand respect pour tous les grands noms cités, mais je m'étonne de l'absence totale de référence aux travaux d'André Martinet depuis le début de ce Congrès, et même dans l'aperçu de M. Rigault hier, qui cite Jakobson, Saussure, Chomsky en oubliant que la description d'Hauteville (1956) a servi d'exemple à bien des travaux ultérieurs.

Je veux rappeler que Martinet a été l'un des premiers à reconnaître et à étudier les changements linguistiques (cf. Economie des changements phonétiques, 1955); il a toujours dit: "Une langue change parce qu'elle fonctionne".

Récemment, il a cultivé et circonscrit la notion de synchronie dynamique qui, différente de la diachronie conçue comme l'étude et la comparaison de deux états de langue et de la synchronie conçue comme constat d'un état de langue, englobe non seulement l'analyse des variantes dans ce même état de langue, mais encore les prédictions de son évolution.

Cette notion de synchronie dynamique me semble intéressante dans le cadre des discussions de ce matière, c'est pourquoi j'ai voulu la présenter. (cf. Evolution des langues et reconstruction, Paris, PUF, 1975)."

Tore Janson (Stockholm): "Language is not only spoken; it is also heard, and the expectations of the hearer must also be changed. So it is important and possible to study the reactions of the hearers, e.g. in experiments with synthetic stimuli. I have done some experiments and would like to get in touch with people working in this area. The results so far are very interesting."

Lars Brink (Copenhagen): "We have tried to show that the forms of a capital city can be spread purely by contagion, according to what we call 'the Napoleon principle': "The enemy is beaten where he is weakest and is immediately enrolled in the

victor's troops." Of course prestige plays a significant role, but not in spreading new pronunciations. The innovations were never felt to be prestigious. Some innovators may be so, but not their followers, and the innovations would therefore drown in traditional forms.

Henning Andersen (Copenhagen): He called for greater precision in the expression of ideas. He did not think Brink said exactly what he meant when he said that a capital like Copenhagen could spread its forms to the countryside. You do not spread changes. It is the people who change their language to conform to the norms as they perceive them in the capital. He then entered a plea against Haugen's view of language as non-systematic or at least finding the non-systematic parts as more interesting. "We won't understand how more or less stray variation that goes on in speech production at all times may become codified and integrated into a system unless we study it in relation to the systems (or the code) that underlie speech production. Labov's study shows that even minute changes are accessible to some degree of subconscious awareness and confirms that what happens when variations turn into a kind of drift is precisely that what could be stray variation becomes a sort of fashion (and here I subscribe to Haugen's view) and is integrated. If we want to explain how changes can be integrated into one system, but not into another, or how changes can occur in one language but not in another, we need to refer to the systems that the stray variations can be integrated into." He then cited Roman Jakobson's opening statement to the Congress, read by Rischel, to the effect that "there is no gross sound matter in language: everything is formed", etc.

Irmgard Mahnken (Saarbrücken): "The question has been raised of how changes can arise in a homogeneous speech community. There are languages which have not changed for a very long time, and others that have been changing and then have stabilized themselves. At least theoretically we need a model of non-change as well as one of change, especially in the development of literary languages. Very little work is being done on the latter, since the social aspects now being investigated are based on living languages. The question of prestige and of social expression can explain many things now under discussion."

Helmut Lüdtke (Kiel): Sound change is predictable. The question is: how and how far? For example, if we knew Latin but no Romance language and wished to predict in what way a Latin word like clave might change in 2000 years, we could choose from the forms written on the left-hand side of the blackboard. Lüdtke suggested that a limited number of possibilities existed, and one would not choose something like akulavic or que. Sound change moves in an irreversible direction, toward shortening. Lüdtke has a theory which he may explain at the next congress. Sound change is reduction: the allegro forms of today are the lento forms of tomorrow.

Eli Fischer-Jørgensen: "I started changing my language when I was fifty and have continued until now. I spoke a conservative form of standard Danish when I was young, and now I find myself using a pronunciation which is approaching what I consider 'vulgar' Danish. This has happened unconsciously and against my will (but the change appears quite clearly from tape recordings). This is quite contrary to some of the ideas presented here." (J. Lund later commented that this might be due to her having a higher linguistic consciousness than most others.)

Richard Coates (Sussex): One often gets the impression that sound change is either community-internal or due to some catastrophic eruption into the community. Coates wished to point out a third mode which has occurred in the literature recently: a new norm external to the community has been integrated into the linguistic system by the adoption of personas by young children. This is exemplified in the work done by Reed in Edinburgh and recently published in the Trudgill volume of readings. Children who were well grounded in the local dialect were able to adopt pronunciation personas taken from TV personalities, disc jockeys, etc. A well-known boxing commentator's mode of presentation was adopted to describe playground fights by particular children. Here is a new norm, a new vector not due to ordinary situational interaction. It is potentially usable independently of the originally appropriate situation. More than one norm is being sanctioned within the system, highlighting once again the dynamic synchrony which has often been mentioned as a feature of these discussions.

Gilbert Puech (Oullins): [In the absence of a written text, the speaker's French is translated into English.] Puech noted

that changes had here been presented as due to social and geographic stratification across a linguistic community. This view should be complemented by studying the need of a social group for a marker of its identity, a change which concerns the weakest point in its system. Therefore he posed this question to Professor Labov: For Philadelphia modifications have been pointed out as due to the lower middle class. Does this correspond to the emergence of this group as a social category which needs to emphasize its identity more strongly by initiating or accelerating linguistic changes? Is it an active or a passive behavior, a consequence of the existing division?

Pierre Léon (Toronto): "(1) Au sujet de la durée des changements -- question posée par Haugen -- certains changements peuvent être très rapides (cf. Léon: L'accent en tant que métaphore sociolinguistique, French Review, 1974). Les ruraux prolétarisés d'un village du centre de la France ont adopté certains traits de prononciation urbaine (parisienne) et prolétaire (ouvriers de la banlieue parisienne), en moins de 10 ans. (2) Ce changement est ce que Léon appelle le résultat d'une conduite idéologique. La nouvelle articulation des ouvriers du village est ce que Birnbaum nomme ici 'a conceptualized (verbalized) mirror image of mental activity' et Fónagy un processus 'métaphorique'. Faudrait-il dire métonymique? (3) Au sujet de savoir qui est responsable de la variation -- question posée par Haugen, Brink, Lund et Labov, Léon donne des exemples des facteurs de la variation dans son village: jeunes, adultes, hommes, prolétarisés. Dans une enquête sur la standardisation des prononciations dialectales de la France (Léon et Léon, à paraître dans les Actes des Congrès de Miami), les facteurs de la variation se groupent en 2 séries oppositives:

$$\text{standardisation} \leftarrow \left\{ \begin{array}{lcl} \text{jeunes} & \neq & \text{vieux} \\ \text{citadins} & \neq & \text{ruraux} \\ \text{mobiles} & \neq & \text{sédentaires} \\ \text{favorisés} & \neq & \text{défavorisés} \end{array} \right\} \rightarrow \text{statu quo dialectal}$$

Tous les facteurs n'ont pas le même poids. (4) Le concept de l'hétérogénéité ordonnée de Labov se retrouve dans les exemples données par Fónagy et se confirment dans les résultats de l'enquête de P. Léon et M. Léon, qui montrent, à côté de la disparition des systèmes de marques dialectales, une diversification au niveau des types de discours. (5) Le concept de sociolinguistique, tel qu'il est employé actuellement n'est-il pas trop restreint aux phénomènes d'indexation des classes sociales, éventuellement aux catégories sexe et âge? Ne faudrait-il pas tenir compte des marqueurs professionnels (Fónagy) et stylistiques dans une approche phono-stylistique plus large (Léon 1971, Essais de Phonostylistique, Didier) tenant compte des facteurs expressifs des situations de communication?"

Anatoly Liberman (Minnesota): On the predictability of sound change he agreed with Haugen: it can always be explained afterward. There are so many things that can happen that given our framework to-day, the framework of system, which is such a very nebulous thing, we can hardly predict what will happen. Also, some things are more probable than others; but given a proto-language and 100 dialects, it is humanly impossible to predict the future. We can only sometimes predict the past, i.e. explain what has happened, but even that is tremendously difficult.

Birnbaum: "I share fully Professor Elbaz's surprise that in all these papers the name of André Martinet was never mentioned". "In a side comment I referred to Martinet's dictum: 'Language is a balanced system with continuous functional redistribution'. To T. Jansson Birnbaum remarked that we all agree that speech perception is important in sound change. Henning Andersen's whole model is related primarily to perception. To I. Mahnken: "Andersen's model was developed to account for historical changes in a Czech dialect." To H. Lüdtke: "I would not call your procedure 'prediction', but educated guesses about probabilities." Reduction is important, but the factor that counters it is the need of explicitness. These forces are constantly in conflict, and it is very difficult to say which will win.

Labov: (1) On women's speech: we do not all agree that it is more advanced. Where women play a part in national life, they are more sensitive to the national prestige, once a sound change has reached maturity and is stereotyped. They are also normally the leaders in linguistic changes from below or unconscious change, where we are hypothesizing a different kind of prestige. (2) This has not been a panel dealing with restraints on linguistic change. However, following Weinreich's paradigm, many

of the sound changes discussed here do show very powerful unidirectional principles, such as the fact that tense vowels always rise. -- On the question of the upper working class: that is not a final characterization of the group involved because it turns out that the role of these innovators in linguistic change is characterized even better by factors having to do with communications research. They are leaders in certain community networks which are very intense locally, but which reach outside the community, and so we get a relatively homogeneous city dialect. Do they emerge as a new group with a need for identification? "I suspect that Professor Puech's characterization was correct. It is not necessarily a new group. It may be an old group that needs to reinforce its identity. These mysterious factors of prestige which we cannot make explicit may be the result of pressures from new groups entering the community. These are challenging the position of the old group. Just as an adolescent must reassert his position in his parents' community, so the Irish or the Italians or the upper working class may be under pressure from Blacks, Puerto Ricans, and other new groups entering the community. Yes, I suspect that the pressure to reassert identity is the driving force behind this continual renewal of sound change."

Suzanne Romaine (Birmingham): Labov's research is an important attempt to deal with the problem of the transmission of change. But the value of the work being done on social factors in sound change is not (as Labov seems to think) to provide explanations of why language changes, but to give us a taxonomy of how social factors interact with linguistic structure in the implementation of language change.

Haugen: "I think we are still in the midst of a very important and very interesting discussion. I thank you for listening to this segment of a discussion that I am sure will go on at future congresses as well as between congresses."

SYMPOSIUM NO. 5: TEMPORAL RELATIONS WITHIN SPEECH UNITS

(see vol. II, p. 241-311)

Moderator: Ilse Lehiste

Panelists: George D. Allen, Robert Bannert, Christopher J. Darwin, Hiroya Fujisaki, Björn Granström, Dennis H. Klatt, and Sieb G. Nooteboom

Chairperson: Claes-Christian Elert

ILSE LEHISTE's INTRODUCTION

The title of the symposium leaves open the question of the type and size of the speech units. The contributors to the symposium have indeed chosen to address themselves to units of quite different types and sizes. Likewise, they have approached the problems connected with the temporal structure of speech units both from the perspective of speech production and from that of speech perception. The contributions include highly theoretical papers, papers presenting detailed results of experiments, and papers falling between these two poles. Some systematization appears to be in order. I would like to present herewith a framework within which I believe the issues can be profitably formulated for the discussions which I hope will follow.

The framework involves three dimensions. One of them concerns the relationship between timing control in production and the role of timing in perception. The second dimension deals with the direction of determination in the temporal organization of spoken language: specifically, with the question whether the timing of an utterance is determined by its syntax, or whether there exist rhythmic principles in production and perception that are at least partly independent of syntax. The third dimension follows directly from the previous two and relates to the type and size of speech units. What is the nature of those units, and are they to be established on the basis of a morphosyntactic analysis of the sentence, or on some kinds of independent phonetic criteria?

Clearly both production and perception are involved in oral communication by spoken language, and it would seem unnecessary to elaborate the point. However, I have had occasion to argue--against considerable weight of opinion--that durational differences in production, be they ever so significant statistically, cannot play a linguistically significant role if they are so small as to

be below the perceptual threshold. It would be wise, I think, to remind oneself periodically of "the evident fact that we speak in order to be heard in order to be understood" (Jakobson et al. 1952). I hope, therefore, that in our discussion of temporal relations within speech units, models of production and models of perception will be related to each other.

The second and third questions concern the direction of determination: does phonology follow syntax, or are we dealing with interacting, but parallel hierarchies? Some researchers have developed programs for generating the temporal structure of a sentence on the basis of segments and syntactic structure, without paying any attention to rhythm. This is, I believe, due to a particular theoretical orientation. Generative phonology operates with segmental features; even suprasegmental features are attached to segments. And in a generative grammar, phonetic output is the last step in the generation of a sentence. An independent rhythm component simply has no place in the theory. For these scholars, then, the speech units are segments, phrases, clauses, and sentences. (And it is quite interesting to see them struggle with units not foreseen in the theory, like syllables and phonetic words.) Researchers who are not fully committed to this theoretical viewpoint operate with certain other units, such as speech measures or metric feet. Again, the reality of both kinds of units can be studied from the point of view of production as well as from that of perception.

Practically all the issues I have outlined are treated in the papers contributed to this symposium. Production is the main concern of the papers of Allen, Bannert, Klatt, and Öhman et al.; perception is the focus in the papers of Carlson et al., Donovan and Darwin, Fujisaki and Higuchi, Huggins, and Nooteboom.

In my brief summary of the papers, I shall address some specific questions to the authors, and raise some general questions that I hope will be discussed at the end of the presentations.

Among the papers dealing with production, Bannert considers the relationship between the durations of vowels and consonants in stressed syllables of disyllabic words in Central Swedish--words of the types stöka (V:C) vs. stöcka (VC:). When sentence accent is added to these words, both segments are lengthened, but by unequal amounts. The increase is largest for the long segment of each type of sequence, i.e. the long vowel in stöka and the long consonant in stöcka. Bannert finds that the temporal structure of quantity is best described by using the concept of vowel-to-sequence ratio, V/(V + C), and he proposes that the VC sequences be viewed as units of production and perception.

I have a comment and a question. The comment relates to the observation that lengthening affects the long segment of the VC sequence. It might be useful to recall here that already Trubetzkoy defined the difference between long and short phonemes in terms of stretchability: tokens of long phonemes are stretchable, while short ones are not. Knowing that it is the long element that is stretchable, one could have predicted Bannert's result: that the addition of sentence accent to quantity increases the temporal distance between the two word types.

The question concerns Bannert's proposal that VC sequences be viewed as units of production and perception. I would like to know how such units relate to already well established units such as syllables. Presumably the syllable boundary falls before the single intervocalic consonant in words like stöka and within the long intervocalic consonant in words like stöcka. I find it difficult to conceptualize the psychological reality of the VC sequence as distinct from segments on the one hand and syllables on the other. It seems to consist of non-comparable parts of the two syllables. Where would these VC sequences fit in a hierarchy of units of production? And what is the evidence for the claim that they also constitute units of perception?

The paper by Klatt presents a detailed scheme for the synthesis by rule of segmental durations in English sentences. It is an almost pure example of that approach that starts from an abstract linguistic description and ends up as a sequence of segments whose durations are conditioned by other segments and by syntactic constraints. The paper does not address itself to the question of overall speech rhythm. A companion paper by Carlson, Granström and Klatt is devoted to testing the output of Klatt's synthesis algorithm. Among the interesting results are the observations that certain aspects of the durational pattern are of greater perceptual importance than others. Vowel duration is more important than consonant duration; the durations between stressed vowel onsets seem to constitute a particularly important aspect of sentence structure. Now it is known that English is a stress-timed

language; there exists an extensive literature dealing with isochrony in English, and some of the arguments in favor of the existence of isochrony are quite persuasive. I would like to address a question to the three authors of the two papers, concerning the role of rhythm in the production and perception of English sentences. Would it not be advisable to include a rhythm component in the synthesis scheme?

The papers by Öhman et al. and by Allen concern themselves with production models in general. Öhman's et al. paper argues for a gesture theory of speech production. The authors claim that "the linguistically functional, intended acoustic effects are not, in general, required to have any particular duration; ...acoustic segments with quasi-stationary qualities will arise not as a final end of the phonetic action but as a secondary consequence of the effort to reach a certain final end (the simultaneous sounding of the effects in question)". Öhman and co-authors maintain that the phonological contrast between Swedish words like vila and villa can be eliminated using this analysis. Namely, the stress effect, which takes relatively long to produce, is coarticulated with the vowel /i/ in vila--thus making the quickly producible /i/ long, while the stress is coarticulated with the sequence /i . l/ in villa, thus making the /l/ long.

I would like to ask the authors--if they were here--how they would handle contrasts between long and short vowels in unstressed position--contrasts which are found in a large number of languages, e.g. in Czech and Hungarian.

Allen's paper draws a useful distinction between descriptive models and theoretical models of speech timing, and makes the intriguing prediction that theoretical models may be about to undergo substantial modification, primarily due to the emergence of an "action theory" of speech production. According to that theory, neural activity is hierarchically organized into successively higher levels of coordination, until the highest level of all can only be described in terms of the overall goal of the action. The models of "intrinsic timing" which Allen describes seem to operate at levels higher than a segment; I would like to ask Allen, too, how the segmental short-long opposition can be handled within these theories. It would have been quite interesting to hear some discussion about the almost diametrically opposed approaches taken

in the papers by Allen and Öhman et al. Öhman, as you may recall, states that manifested segmental durations are generally secondary consequences of the effort to produce simultaneous acoustic effects. Thus there appears to be no room for temporal programming as such. The models Allen refers to claim that intrinsic timing is an inherent property of the speech act. Can these two views be reconciled, or will one of them be proved wrong?

Among the papers devoted primarily to perception, Nooteboom's presents a decision strategy for the disambiguation of vowel length in Dutch. The strategy presupposes knowledge on the part of the listeners of temporal regularities of speech, and the ability to shift an internal criterion--the boundary between long and short vowels--depending on the speech context. For example, the listener is assumed to know that vowels followed by pause are generally longer than vowels followed by a consonant; that vowels are longer when that consonant is a fricative than when the consonant is a plosive; that vowels are shorter with increasing number of unstressed syllables following the syllable containing the stressed vowels, etc. Nooteboom hypothesizes that listeners do indeed possess this knowledge and shift the perceptual boundary between long and short vowels according to speech context. The data presented by Nooteboom are quite impressive; it seems to me, however, that there is something artificial in the described situation. When the listeners adjust the criterion depending on the speech context, they are in fact perceiving the total speech act, not just the vowels. Otherwise there would be no need to perform the adjustment. The environment is just as much part of the percept as the vowel. From my experience with English, I would predict that the durations of vowels and postvocalic consonants stand in a compensatory relationship, and that both are related to the overall duration of the word. Even though the strategy Nooteboom proposes is quite complex, I submit that it is actually an oversimplification.

Fujisaki and Higuchi present an analysis of the temporal organization of segmental features in Japanese disyllables consisting only of vowels, and find that although the onsets of the transition for the second vowel are distributed over a relatively wide range, a perceptual analysis of the onset of the second vowel shows relatively little temporal variation. It thus seems that the apparent diversity of the onset of transition in various disyllables

is introduced for the purpose of maintaining the uniformity of perceived duration of segments. Fujisaki and Higuchi consider their results supportive of a model in which the motor commands and the articulatory/acoustic realizations of successive segments are programmed in such a way that the perceptual onsets of successive segments are isochronous.

I am quite impressed and convinced by these results and would really like to have more information. Japanese and English appear to have quite different temporal structures at the sentence level. How far does isochrony go in Japanese? Is the disyllabic sequence conceivably a basic unit of temporal programming--for example, if we have a word of four syllables, does it have the length of two disyllabic sequences? Is there any interaction between segments and syllables--for example, how would the inclusion of consonants in the disyllabic sequences influence their duration both in production and perception?

The paper by Huggins is mainly concerned with the intelligibility of temporally distorted speech. Huggins finds that a distorted timing pattern (which often characterizes the speech of the deaf) is a sufficient cause for catastrophic loss of intelligibility. While I have no argument with this particular claim, I would like to take issue with a statement concerning the relationship between pauses and other cues employed to indicate syntactic boundaries. Huggins states that boundaries that are marked by pauses need not be inferred from more subtle cues. In some recent work of mine on the perception of sentence boundaries, I found that listeners can completely ignore a fairly lengthy pause, if it is not preceded by a certain amount of preboundary lengthening and/or change in fundamental frequency. I wonder if Huggins would really persist in claiming that pause is a sufficient boundary signal?

The paper by Donovan and Darwin deals with the perceived rhythm of speech, with special consideration of the problem of isochrony. Their paper tests, among others, a hypothesis that I had formulated in 1973 and discussed in more detail in 1977. My observation was that listeners tend to hear utterances as more isochronous than they really are, and that listeners perform better in perceiving actual durational differences in non-speech as compared to speech. I concluded from this that isochrony is largely a perceptual phenomenon. Donovan and Darwin have confirmed

these results. They make two points in addition: first, that isochrony is a perceptual phenomenon which is not independent of intonation, and second, that it is a perceptual phenomenon confined to language, reflecting underlying processes in speech production. Donovan and Darwin question the value of seeking direct links between syntax and segmental durations rather than indirect ones by way of an overall rhythmic structure.

While I am in enthusiastic agreement with this particular conclusion, I would like to question the presumed role of intonation in establishing the rhythm of spoken language. There is recent evidence (De Rooij 1979) that intonation contributes very little, if at all, to the temporal structure of a sentence: perception of the temporal structure is not noticeably changed when the fundamental frequency is changed to a monotone. In some unpublished work I found that syntactically ambiguous sentences could not be disambiguated by manipulation of the fundamental frequency, whereas they could be successfully disambiguated by systematic changes in the time dimension. (This latter result has appeared in print: Lehiste, Olive and Streeter, 1976.) If Donovan and Darwin persist in their claim, I would like to hear stronger arguments than have been presented in their paper.

The discussion will be structured as follows. The authors will now have approximately five minutes each to make corrections and additions to their papers. Then we will have a panel discussion, lasting about 30 minutes, during which I hope the authors will respond to some of the questions I have brought up--as well as contribute questions of their own that we will all discuss. The last hour of the session will be devoted to a general discussion with participation from the floor. If there is time, I shall try to verbalize some of the final conclusions that emerge from the discussion.

References

Jakobson, R., C.G.M. Fant, and M. Halle (1952): Preliminaries to speech analysis, Cambridge, Mass.: MIT Press (tenth printing 1972).

Lehiste, I. (1973): "Rhythmic units and syntactic units in production and perception", JASA 54, 1228-1234.

Lehiste, I. (1977): "Isochrony reconsidered", JPh 5, 253-263.

Lehiste, I., J.P. Olive, and L.A. Streeter (1976): Role of duration in disambiguating syntactically ambiguous sentences", JASA 60, 1199-1202.

De Rooij, J.J. (1979): Speech punctuation. An acoustic and perceptual study of some aspects of speech prosody in Dutch, Dissertation, Utrecht.

COMMENTS FROM THE PANELISTS

[Since it is impossible to reproduce here the slides shown by several of the discussants, those parts of their presentations that refer to slides have been edited to make them reasonably comprehensible without visual aids.]

R. Bannert reiterated his conviction that the domain of quantity patterns in Standard Swedish and in a number of other languages is the stressed vowel and the following consonant, and questioned the claim that the syllable boundary falls in the middle of a long consonant. He also presented additional evidence concerning the effect of sentence accent on the durational structure of words like stöka and stöcka. Sentence accent lengthens not only the durations of the segments which make up the sequences, but it lengthens all segments of the word in focus, including the second, unstressed vowel of the test word. The segments /s/, /t/ and /a/ have the same duration in both types of test word. The clear difference between the two minimally contrastive words is in the VC sequences of complementary length. The significance of the VC sequences has also been confirmed by perceptual experiments.

D. H. Klatt formulated some general questions that relate to the problem discussed in his paper: 1) what are the phenomena to be described in a particular language, 2) how do all the rules interact, 3) what is an appropriate underlying representation for an utterance in a particular language, if one wants to predict durations or do a complete synthesis by rule? In a linguistics framework, one would like to start with an as abstract--but psychologically real--representation as possible. As regards the rhythm component, it is true that the paper makes the impression that no attempt has been made to account for it; but there are some rules that make the segmental patterns tend to be isochronous, such as cluster shortening rules and polysyllabic shortening rules within words (but not within feet). These two rules, and perhaps some interactions of other rules, bring about a tendency toward isochrony.

B. Granström pointed out that the primary aim of their paper was not to evaluate Klatt's rule system, but to look into what things are important in rule systems in general, and how naturalness of a rhythmic structure is related to intelligibility. Isochrony in perception is obviously there, or the observation would not have been made in the first place; the question is how important it is in production. It might be that it is not even desirable to have isochrony in production. Parallel studies of rhythm in music indicate that music generated by computer with perfect isochrony is often very dull. Another reason why we believe isochrony is not necessary in the description of durational structure is that it turned out that the rule system is actually very good: in the evaluation process, the utterances generated by the rule system were evaluated as being more natural than the actual productions by Dennis Klatt! And measurements show that the output of the rule system was more isochronous than the actual productions. We believe therefore that an isochrony component is not needed, at least not for the generation of the types of isolated sentences produced in our experiment.

G. D. Allen asked how one should handle short and long quantity in intrinsic timing models. According to the extrinsic view, the motor plan includes temporal features which are used by an extrinsic controller (a "speech clock"), which somehow signals the motor system when to begin and end a specified activity. In the intrinsic view, however, the temporal properties of the act are never specified as such but rather are the result of other, not specifically temporal properties of the act. As an example, consider long versus short vowels. An extrinsic timing model would deliver the command to produce the segment (e.g. /a/) along with a "start" command and a durational feature, which would be used by the clock to generate a "stop" command. An intrinsic timing model, on the other hand, would select either the short or long /a/, which must be represented as distinct acts within the motor repertoire, and that short or long /a/ would then be produced as an integrated part of the overall syllable, word, and/or phrase. Its resulting duration would be a complex function of the several interacting levels of structure and behavior which all together define the act.

Asking how one might test for the existence of intrinsic versus extrinsic timing, Allen reviewed an experiment by Laver (cf. J. Laver's comment below) as an example of a potentially useful experimental paradigm.

S. G. Nooteboom presented some data showing that the perceived boundary between short and long vowels shifts in accordance with speech production regularities. The listener has at his disposal a very detailed knowledge of the temporal regularities of speech: he knows how speech should sound in his language. It is more difficult to know how the listener uses this knowledge, and even more difficult to know how it is stored. In the paper, Nooteboom had made a proposition that all this knowledge is stored as a set of rules in the brain, and that the listener rapidly calculates the expected durations of both short and long vowels, places his criterion in the middle between these two, and thus adjusts his judgment according to context. He considers this now to be a very unlikely procedure, mainly because it must be time-consuming to do so much calculation, and also because he does not believe that all these higher-order effects are going straight back to the level of phoneme decision. There is another way of accounting for the same data, in accordance with some psychological models of word recognition.

H. Fujisaki stated that the motivation for his contribution to this symposium was to provide some quantitative means and frameworks for discussing temporal relations within speech units. The successive units of connected speech manifest themselves not as discrete, separable acoustic events, but rather as overlapping and mutually interfering events. Thus, for example, in discussing the issue of isochrony, one cannot claim that a certain point represents the timing of a speech unit just by looking at the speech signal waveform or its spectrogram. In order to decide whether isochrony is a characteristic of speech production or of speech perception, experimental techniques are needed that allow one to infer the timing of the production of segments as well as the timing of their perception. In his paper, Fujisaki showed quantitative techniques to determine these timing relationships. Thus his contribution was concerned not only with perception, but also with production. The material was deliberately restricted to disyllabic two-mora words of Japanese, since they can be regarded as the smallest examples of connected speech. The materials were further restricted to disyllabic words consisting only of vowels (which are quite common in Japanese), since the articulatory transition from a vowel to the following vowel can be most clearly observed and analyzed from the trajectory of formant frequencies.

Presenting several slides to illustrate the points made in the paper, Fujisaki pointed out a rather wide range of distribution of the onset of articulatory transition among utterances with different combination and order of vowels. At the same time, a strong negative correlation was found between the onset time of such a transition and the rate of transition. In other words, slower transitions were almost always initiated earlier, while faster transitions were almost always initiated later. The onset was distributed over the range from 90 msec to 150 msec within a total utterance duration of approximately 300 msec, which is at least several times larger than the DL for the perception of temporal differences at these durations.

The determination of perceptual timing is based on listening experiments using the same speech material, but by truncating the waveform at various points and presenting only the initial portions as stimuli. The time instant corresponding to 50% judgments was defined as "the perceptual onset" of the second vowel (syllable). The perception of the second vowel starts not at the onset of the formant transition, but at some point where more than 60-70% of the total formant transition has been traversed. The perceptual onsets of the second vowel in various disyllables are concentrated within a very narrow range (about $\pm$ one DL) centered around the midpoint of the utterance. Thus the initial and the final vowels are almost always perceived as being of equal duration within a vowel disyllable. The results indicate that the isochrony in this case is neither a mere illusion nor a perceptual distortion of the acoustic reality, but the timing of perception actually occurs isochronously. These findings may be interpreted in the light of a model for the control of speech timing (cf. Figure 7, p. 281 of Volume II). One may safely assume that the articulatory control under ordinary utterance conditions is open-loop control. The findings of this research support the hypothesis that motor commands are programmed in such a way that the perceptual durations of the two vowels within a disyllable are perceived as equal, at least as far as Japanese vowel disyllables are concerned.

In reply to Lehiste's questions, Fujisaki remarked that the work is presently being extended into two directions. One is the case of sequences of three or more vowels which are also quite common in Japanese. Preliminary results indicate that the same conclusion holds for these polysyllabic words. The other direction for future study is to include CV-syllables. It is necessary, however, either to establish an analysis technique whereby one can infer from the speech signal the exact timing of consonantal articulation, not just its acoustic consequences, or to rely on physiological observation to determine the timing of speech production and compare it with the timing of speech perception.

C. J. Darwin recalled the purpose of the reported experiment: to distinguish perceptually between two models for the production of speech durations. According to one model, each phoneme has a sort of "platonic" duration which is shortened as a function of syntactic influences; according to the other, there is an underlying rhythmic structure which is perturbed on the basis of the incompressibility of the elements that one is trying to fit into it. The prediction from this theory is that we are aware of the underlying regular rhythmic foot rather than its surface manifestation.

Darwin also presented additional data which supported the claims made in the paper--that people perceive rhythm to be more isochronous than it really is, and also that this does not apply to non-speech. Additional work has been done at Sussex addressing the question whether syntactic boundaries are signalled just by phrase-final lengthening or by lengthening the whole foot in which the boundary occurs. The results show that the latter is the case.

DISCUSSION

I. Lehiste recalled the results of some of her earlier experiments which had shown that speakers can use several strategies to signal syntactic boundaries. The strategies have a common result, namely lengthening the foot containing the boundary. These experiments had not tested the relative importance of the different strategies, e.g. of phrase-final lengthening, as boundary cues. Lehiste challenged Klatt and Granström to respond to Darwin. In the discussion which followed, it emerged that even though lengthening of the foot is of primary importance, it does matter what part of the interstress level is lengthened: listeners feel uncomfortable if the lengthening is limited to the part that follows the syntactic boundary. It appears that both phrase-final lengthening and lengthening of the foot are necessary for listeners to identify the position of a syntactic boundary.

G. D. Allen commented that it is perhaps wrong to call isochrony in English "largely perceptual" (as had been done by Lehiste), since speech is already temporally highly structured in production. He also questioned those of Darwin's results that showed that nonspeech was not perceived as more isochronous than the stimuli really were. This finding appears to be at variance with previous research on time perception, and Allen therefore asked (1) was there in fact a trend in the right direction which was smaller than the one for speech and not statistically significant, and (2) what would be the effect on the nonspeech temporal interval perceptions of filling the intervals with various sounds, as the intervals of speech are filled?

C. J. Darwin responded saying that one of the nonspeech results did depart significantly from actual durations, but it went in the other direction--it was perceived as significantly less isochronous. Darwin agreed with the need to perform experiments with different kinds of nonspeech controls with filled intervals. He would also like to perform similar experiments with music.

I. Lehiste expressed the hope that temporal patterning in other languages besides English and the Scandinavian languages might be considered during the discussion, and urged the discussants to remain conscious of the general theme of the symposium: what are the units within which temporal structures are manifested, how does sentence rhythm relate to the durations of these smaller units, and how does sentence rhythm relate to nonphonological aspects of language--e.g. to syntax.

B. Granström found that perhaps too much attention had been given to isochrony in the discussion, and presented some data that showed that a word can be a very important unit for temporal programming.

P. L. Divenyi, referring to his 1977 dissertation, stated that he had found context effects in rhythmic perception in music. If there is no isochrony in the microscopic sense, there could be in the macroscopic sense, even for nonspeech. Rate is a variable that can affect rhythmic perception. Isochrony is an inherent

property of the production system; one could relate isochrony found in perception to production by simply postulating certain listening habits. Thus he does not see any contradiction between productive isochrony and perceptual patterns found in perceptual experiments.

L. Lisker suggested an experiment: to assign segment durations by a random process (in synthesis), and find out what loss in intelligibility and naturalness there would be.

R. Gsell discussed temporal relations in Thai, a quantity and tone language. Stress has a leveling effect on quantity contrasts. Temporal constraints and perceptual limitations produce for the listener neutralization of contour tones in shortened and unstressed syllables.

E. Selkirk took issue with the moderator's characterization of generative phonology as a theory which is in principle unable to countenance such notions as syllable, timing, and rhythm. The notion of the phonological representation within the theory was one of a purely linear kind which saw it as a sequence of segments and boundary elements. In recent years, though, workers who see themselves as operating within the context of generative phonology have been rediscovering that this conception of phonological representation has to be radically revised, allowing for far richer hierarchically arranged suprasegmental structures.

Some workers, Selkirk included, have been arguing for a rather different conception than that in the Sound Pattern of English of Chomsky and Halle, of the relation between phonology and syntax in a generative grammar. In this conception, syntax is seen as bearing on phonology only insofar as phonological units, like syllables or intonational phrases, may have specific syntactic domains over which they are defined, but phonological and phonetic processes are seen as functioning only in terms of these phonological hierarchical structures. It is a claim of this theory that something like final lengthening has its domain defined in terms of phonological units (such as intonational phrase and perhaps others); it would not be immediately sensitive to syntactic structure. What is predicted here is that there would be a systematic convergence of various types of phonological phenomena; the unit at the end of which one finds lengthening would be the same one with which, for example, an intonation contour would be associated, or it may also be the domain of rules of segmental phonology.

Lengthening or the realization of intonational contours and so on are not conceived as individually and separately sensitive to units of syntactic structure.

H. Fujisaki, responding to comments by P. Divenyi and L. Lisker, agreed that we need to look at both microscopic and macroscopic levels of timing. There should be a hierarchy of levels in which speech timing is programmed and maintained. For instance, the problem of compensation between the duration of a consonant and the following vowel is a matter of timing within a syllable, but the compensation between the duration of a vowel in a CV syllable and the following consonant of the next syllable is a matter of interaction between sub-syllabic units across syllable boundaries. Fujisaki had looked at vowel disyllables in order to investigate the relationship between durations of the two syllables without having to consider the problem of consonant-vowel compensation.

J. Laver reviewed his "motoric balance point" experiment mentioned by Allen in connection with two opposing views of the nature of the control of temporal relations. The argument is between the extrinsic view of temporal control, where a "speech clock" acts as an external, overlaid control device, versus the intrinsic view, where temporal relations are the direct product of characteristics of segmental representations themselves. Laver singled out one finding in his experiment which tends to support one of these views. When his subjects were faced with the need to produce forms which had a quantity difference as well as a quality difference between them, such as PEEP and PIP, then the link between quantity and quality was very labile in their productions, and very easily perturbed. There were many errors made, where the right quality but the wrong quantity was produced. So there were examples of PIP with a long vowel duration and of PEEP with a short vowel duration, where both nevertheless showed appropriate articulatory quality. This tends to support the extrinsic view, where duration is at least to some extent the product of specific neuromuscular programming separate from programming for articulatory spatial targets as such.

N. Thorsen addressed a question to Nooteboom, who, with his last slide, had appealed to the audience to have the courage to assume that word identification precedes phoneme recognition. Thorsen asked how Nooteboom would account for the perception of

slips of the tongue, which are generally perceived as such, i.e., as slips or mistakes, while at the same time the word is being identified correctly.

K. L. Pike, in his comments, made the point that in English, both isochronic and non-isochronic timing are essential. Under certain circumstances, we must not have isochronic stress groups; under other instances we must indeed have them. This is connected with the fact that in his normal use of English there are some items which one might call "double stresses". These are, in general, related to certain kinds of syntactic groups. There is also a kind of a semantic component which often goes with these double stresses. It is a unitizing effect, tying the items together in some kind of a single concept to be viewed as a unit rather than as components loosely strung together. We must not be so inflexible that we assume that we must have either isochronic stress groups or else we must have largely non-isochronic stress groups. In Pike's analysis of the material one must leave room for both in English. This, in its turn, forces another conclusion: we cannot assume that there is a single rigid set of rules mapping directly, and in only one manner, material from the grammatical hierarchy on to the phonological one; nor of semantically oriented units from a referential hierarchy on to the grammatical or phonological one. We need three hierarchies, always interacting one with another, but never the one totally determining the other. Our rule systems, therefore, cannot be inflexibly from grammar to semantics and phonology; nor from semantics to grammar and then phonology. Rather we must have some interdependence in which the purpose of the speaker is distributed in ways which are vastly more complex than a one-way rule system can tell us.

S. Nooteboom, responding to Thorsen, disclaimed having ever implied that listeners cannot extract phonemes from the acoustic signal. In the normal recognition of known errorless words--which is usually very fast indeed--it is not necessary to assume that phonemes are mediating in perception. Hearing unknown words, or words containing detected mispronunciations, listeners must have been listening in a "phoneme mode".

L. Nakatani questioned the existence of isochrony in production. Even though in comparing black dog with blackish dog there seems

to be isochrony, this can very easily be explained by the fact that the first syllable in a bisyllabic word becomes shortened relative to the same syllable in monosyllabic words. There is another factor operating here--resyllabification. In blackish, the /k/ is aspirated, indicating that the /k/ now belongs to the second syllable. So one cannot compare black and blackish, for the syllables are different. If one controls for this by using reiterant speech, some kind of compensation can indeed be found; but if one controls for that and looks at the effect due to the insertion of an unstressed syllable in medial position, one does not find any compensation. Similarly, if one inserts an unstressed syllable at the beginning of the second word, there is no compensation. There is a very linear relationship between the number of intervening unstressed syllables and the interval between stressed syllables. This is consistent with data collected by Wayne Lea.

Nakatani has also looked at duration patterns of words in different contexts. If there is a tendency toward isochrony, the durations of words should vary as a function of the context in which they occur. Looking at the same words in different positions in different sentences, Nakatani found that the duration patterns of words were extremely consistent, and concluded that there is no evidence for isochrony in production. Therefore it should be ascribed primarily to perception, and be based on the fact that content words and function words alternate, and that most bisyllabic words in English have the stress on the initial syllable.

I. Lehiste remarked that there are usually several principles operating at the same time, and they interact. Tendency toward isochrony is one of these principles, but there is certainly another one--the principle of maintaining the temporal integrity of the word, so that the duration of a monosyllabic word is roughly comparable to the duration of a disyllabic word. When these two principles interact, they will influence each other.

E. Uldall noted that we are devoting our attention almost entirely to "stress-timed" languages (though there have been references to Japanese). She expressed the wish to hear a lot more about the opposite case: for example, about French. Phoneticians very frequently refer to English as a classic case of stress-timing, and to French as a classic case of syllable-timing. Yet all the experimental evidence we have about English shows that the

"rhythmic feet" are far from isochronous, and what Uldall has seen of French syllables makes her think that they are not isochronous either. So why do phoneticians go on saying what they do?

G. Fant stated that most of our data about durations have been obtained from speech waves--oscillograms and spectrograms. The question is, can we interpret this in terms of a production model to give a better perspective? The answer is affirmative. For instance, if we study vowels in sentence-final stressed position, we find that all the durations are the same, because what has determined the termination of the vowel is the phonatory gesture which is the same for all vowels and independent of the preceding consonant. On the other hand, if the vowel is followed by a consonant, the consonantal frame influences the vowel duration. This is the articulatory aspect. So the duration of a vowel can be set either by phonation or by articulation or, really, both. If a voiced stop comes after the vowel, then of course the vowel is terminated as the acoustical consequence of the constriction, but if it is an unvoiced plosive which comes after the vowel, then there is a separate neural command for the abduction of the vocal cords. That command is somewhat time-locked to articulation, but they are still separate events. This can be a fruitful way of scrutinizing the durational data.

S. M. Marcus gave a brief summary of his research concerning Perceptual Centres or P-centres, which involve rather more fine-grain aspects of speech timing than those determining the temporal structure, isochronous or otherwise, of continuous speech. In producing perceptually isochronous sequences of isolated monosyllables, perceptual regularity corresponded to no simple physical alignment. Subsequent experiments have shown the P-centre locations to be a function of the acoustic structure of the whole stimulus--for example extending the /t/ closure of "eight" shifts its P-centre. These results clearly demonstrate that before considering such questions as isochrony and "syllable-" or "stress-timing" in continuous speech, we need to be very clear what we are measuring the timing of. We must be wary of assuming that simple instrumental measurements, such as consonant and vowel onsets and durations, are related in other than a complex way to our perception. We should also be aware that much of the data which has been used to demonstrate either isochrony or lack of isochrony now needs to be carefully reexamined.

G. D. Allen urged the audience to view timing and rhythm as mental phenomena. Time as it is measured in spectrograms and oscillograms is but one correlate of timing and rhythm. These phenomena belong in the mind, several levels removed from the articulatory periphery.

I. Lehiste thanked the panelists, the very efficient chairman, and all contributors from the floor. She observed that many issues had remained unsolved--for example, the question whether isochrony in English is a property of production or perception. One underlying assumption, however, appears to have been generally accepted--namely that temporal organization operates within units that are larger than a single segment. The task still remains to establish these units for different languages. She concluded with the hope that this discussion has contributed some background that will be taken into account in future research directed toward the discovery of the temporal structure of language.

SYMPOSIUM NO. 6: MOTOR CONTROL OF SPEECH GESTURES

(see vol. II, p. 315-371)

Moderator: James Lubker

Panelists: R.A.W. Bladon, R.G. Daniloff, Hajime Hirose, Peter F. MacNeilage, and Joseph Perkell

Chairperson: Leigh Lisker

JAMES LUBKER's INTRODUCTION

In preparing my introductory comments for this symposium I have made two assumptions: first, I am assuming that those of you in attendance are interested in speech production/motor control theory and have therefore taken the time to at least glance through the papers for this symposium as they were published in volume II; and secondly, I am assuming the goals of phonetics to be as described by Björn Lindblom in his plenary lecture (p. 3-18, this volume).

Acceptance of the first of these assumptions implies that I need not spend much time in summary of the papers in this symposium; they are there for the reading. Rather, I will take as my goal to provide a common framework for those papers and the points of view expressed in them, in order to allow the discussion of current and important issues in production/motor control theory.

Since acceptance of the second assumption will dictate the nature of the framework and issues which we will develop for discussion, it is perhaps wise for me to be somewhat more explicit about it. In the summary (vol. I, p. 3-4) Lindblom states: "Phoneticians accordingly construe their task of speech sound specification as a physiologically and psychologically realistic modeling of the entire chain of speech behavior." And he then goes on to pose the questions of (1) why it should not be possible for "phoneticians to extend their inquiry into the sounds of human speech to ever deeper physiological and psychological levels using speech as a window to the brain and mind of the learner, talker and listener?", and (2) "Why we should not expect more complete, theoretical models and computer simulations to be proposed for speech production, speech understanding and speech development that match the present quantitative theory of speech acoustics in rigor and explanatory adequacy?".

Indeed, the very title of this symposium, The Motor Control

of Speech Gestures, suggests research and theory devoted to an attempt to elucidate the rules and systems "at ever deeper physiological and psychological levels", by which man generates speech, and to do so with as much precision and scientific rigor as possible. Motor control research and theory must be integral to the goals stated by Lindblom, that is, to the development of explanans principles in phonetic and linguistic theory. Thus, the acceptance of those goals is my second assumption for this symposium.

There remains, however, much room for discussion since the search for precise and valid explanans principles for the generation of human speech is currently faced with several crucial issues, which are well illustrated by the papers presented in this symposium. Those issues can be discussed within three very broad and highly interrelated areas of theory and research.

In the first place, many questions in motor control/production research have quite naturally dealt with the form and function of the system or systems which operate to produce a speech acoustic signal. That is, a major effort in motor control research has been the attempt to discover the rules which explain and predict the transformations at the several interfaces in the chain of language generation and perception. Armed with such rules we would indeed have "a window to the brain". And since that is precisely where language resides, knowledge of these rule systems would provide us with a strong tool for the elucidation of certain aspects of language theory. Efforts to discover the rules have not, thus far at least, resulted in a Motor Control version of the Acoustic Theory of Speech Production, but as Lindblom suggests, there is no reason to believe that we will not one day have such a theory. Every paper in this symposium deals via proposed models, specific data or both with the form and content of such rule systems and it would thus seem obvious that this should be a fruitful area for discussion.

A second broad area of theory and research in the motor control of speech gestures is the precise form or nature of the units which serve as input to the motor control systems. In the papers of this symposium a number of possibilities are suggested: Abbs uses a matrix of phonetic features; in an updated version of their paper Daniloff and Tatham also suggest such a matrix. Bladon

considers several possibilities including features, phonemes and phonological syllables; Gay and Turvey seem to be viewing the input as phonemic; Perkell agrees that studies of motor control mechanisms are closely related to the nature of the "fundamental units underlying the programming of speech production", but he does not speculate in this paper as to what those units might be. Although the papers of Folkins, Hirose, and Sussman are concerned with specific experimentation with the functioning of the motor control systems, irrespective of the input unit, the nature of that unit would clearly seem to be a second broad area for useful discussion.

Finally, let me propose a third general area for discussion; an area which is so related and intertwined with the preceding two as to be virtually inseparable from them. It concerns more the form of attack upon the problems of the preceding two areas.

I have been implying that motor control rules of some kind are necessary in order to move from abstract linguistic concepts such as the phoneme or syllable to the concrete data obtained in speech production experimentation. These two sets of units, the abstract concepts of linguistics and the hard data of production research have never been very well matched and if they are to be used together in attempts to explain speech and language generation then transformation rules would, in fact, seem necessary. Fowler et al (1978) have called such efforts "Translation Theories" and they contend that virtually all production research to date may be classed as one or another type of translation theory. Fowler et al also suggest that all abstract linguistic units possess three properties: they are descrete, static, and context-free; while all units of production are dynamic, continuous and context-adjusted. A clear mis-match! Most of us would agree with Liberman and Studdert-Kennedy (1978) that translation from discrete, static and context-free to dynamic, continuous and context-adjusted requires a "drastic restructuring" of segments, whatever the original input segments might be. Thus, the many attempts to provide theories which explain and solve the non-isomorphism between the abstract linguistic units and the concrete production units. In the course of that work much effort has been expended toward attempts to find physical/physiological correlates of the abstract linguistic units.... to eliminate the non-isomorphism.

To date this research has been notorious for its lack of success and physical/physiological correlates of abstract linguistic units are conspicuous largely via their absence. Such repeated failures have caused some researchers to become disenchanted with the particular research strategy entailed in translation theories. They contend that when experimental data are shown repeatedly to be at variance with theoretical constructs it is only natural to begin to question the legality of the constructs. Carried on, such an argument raises the question: should production/motor control theorists develop their own units and concepts which are based on actual experimental observations of motor control mechanisms in general and which are unbiased by notions and abstract concepts borrowed from linguistic theory? Moll, Zimmerman and Smith (1977) have presented perhaps the most explicit and extreme version of this view and they suggest that: "Such an approach might lead us to the identification of units of programming based on the physiological parameters of movement, muscle contractions and neural activity, units which might or might not correspond to any construct previously defined."

Although such a view may be compelling, it can lead to a small feeling of scientific schizophrenia in those of us who have for so long followed the "translation theory road". The notion of sets of transformation rules between such interfaces as the output of a phonological component and the neurophysiological structures of the speech producing mechanism seems such a reasonable notion. The linguistic concept of "phoneme", for example, is indeed an abstract one.... unseen and unseeable. But so also are many of the concepts of the physicist unseen and unseeable. Further, Fromkin and others both previously, and here at this Congress, have discussed persuasively the psychological reality of linguistic units as demonstrated by, for example, speech errors. Nevertheless, the arguments proposed for not allowing ourselves to be prejudiced by the use of preconceived and abstract linguistic notions may also be persuasive and there may thus be some benefit in discussion of this issue.

In any case, we see two quite differing points of view concerning the theoretical and experimental approach to the general problem areas of input units and motor control rules and systems. And, there is yet a third point of view. Berstein's Action Theory

(1967) was originally proposed as a general theory of coordinated movement. Turvey (1977) and his associates (e.g., Turvey et al, 1978; Fowler et al, 1978) have applied this theory to the generation of speech and language. The action theory point of view also argues against the use of translation theories in speech production/motor control research, but does not agree that such research should be conducted without reference to linguistic units. These investigators' use of action theory and their development of such concepts as "coordinate structures" in speech motor control represent an attempt to avoid translation theories while at the same time not rejecting out of hand the use of all traditional linguistic concepts.

And so, the problems regarding our experimental approach to the nature of the input units and the motor control rules and systems which act upon those units would seem to be: (1) Should production/motor control theorists continue to search for translation rules which mediate between abstract linguistic units and concrete production units, or (2) Should production/motor control theorists attempt to ask questions about fine motor behavior in general in an attempt to elucidate speech and language generation and in the process create new or substantiate old input units, or (3) Should production/motor control theorists follow the entirely new course proposed by Action Theory and its claim of understanding linguistic organization via experimental study of the lower, "basic" properties of speech acts without the use of translation rules? I should add, since there was some misunderstanding at the symposium, that I have here only stated these as experimental approaches worthy of discussion and I have not aligned myself with any of them in this paper.

It seems to me that this symposium offers a reasonable forum for the discussion of these very important issues.

Here, then, are three very broad and interrelated areas of research and theory from which we might profitably draw questions for discussion: (1) the nature of the programming units; (2) the form and structure of the system or systems which act upon those units; and (3) what the best theoretical approach might be to discover what those units and systems are.

Each of the papers in this symposium takes up issues in one or more of these broad areas and it may now be appropriate to

consider some of their specific points of view.

For example, one topic which may be of general interest to all of the papers and which may involve each of the three areas discussed above is: What is the nature and the relative roles of feedback mechanisms versus central programming/simulation loops in motor control systems?

In that framework Abbs presents a model which stresses that not only is afferent feedback required in speech control, but it must take place at a variety of sites, including rather low level ones, in order to account for speakers' ability to compensate rapidly to unanticipated disturbances in ongoing speech. While he does not reject out of hand the possibility of a pre-adjustment, or efferent copy, system he argues that afferent control capability is the prime factor in accounting for rapid adjustments to dynamic unanticipated loads.

Perkell, on the other hand, argues that both orosensory feedback and central programming with internal feedback play important roles in motor control. Specifically, he implies a major role for central programming and internal feedback (feedback entirely internal to the central nervous system) "for the moment-to-moment (context-dependent) programming of rapid movement sequences".

Gay and Turvey present still a third possibility in the form of data which they interpret as being negative to the existence of an open-loop control system and positive to the function of the coordinate structures of Action Theory. Their principle argument against any closed loop system, "internal" or otherwise, is that "while an error signal can index how near the collective action of a number of muscles is to the desired consequence, it does not prescribe in any straightforward way how the individual muscles are said to be adjusted to give a closer approximation to the referent."

Several of the papers present data which are relevant to these theoretical observations. For example, in one experiment Folkins provides an indication of the variability, and thus the trade-off in muscle function, for jaw elevation, thereby supporting MacNeilage's (1970) earlier views on the variability of muscle activity for the attainment of particular vocal tract targets. Additionally Folkins shows that the medial pterygoid muscle contracts in a similar manner with or without a bite block in place thus

suggesting that "unnecessary" jaw closing activity is not eliminated either in the equations of constraint proposed by Action Theory or in the central movement plan of a simulation loop.

Data supportive of intermediate stages of feedback control as well as different patterns of control, which tends to support the model proposed by Abbs are presented by Hirose in his study of electromyographic activity and movement of the soft palate.

Sussman's elegant single-motor unit work demonstrates evidence for cellular level reorganization of muscle function in jaw elevation in response to a "behavioral and biomechanical aspect of the encoding program for speech.

These and additional experimental data provided by Folkins, by Hirose and by Sussman must be considered in the theoretical interpretations provided by Abbs, by Gay and Turvey and by Perkell. Perhaps in doing so, and in discussing additional data, we can make some progress in the question of the nature and relative roles of feedback and central programming. Unfortunately it must be noted, in retrospect, that such a discussion was difficult for the panel to initiate, largely due to the fact that several of the authors were unable to attend the congress. Specifically, Abbs, Folkins, Gay, Turvey and Sussman were not present on the panel. Sussman was ably represented by Peter MacNeilage but it was not possible to get the viewpoints of the others in the form of direct discussion.

Nevertheless, with all of these issues, ranging from the relative merits of translation theory versus action theory versus (for want of a better term) exclusively neurophysiologically based theory to the issues of the relative importance of feedback versus central programming, I think that without any more preambling on my part we have more than enough conflict with which to begin a discussion of the motor control of speech gestures.

References

Bernstein, N. (1967): The coordination and regulation of movements, London: Pergamon Press.

Fowler, C.A., P. Rubin, R.E. Remez, and M.T. Turvey (1978): "Implications for speech production of a general theory of action", in Language production, B. Butterworth (ed.), New York: Academic Press.

Liberman, A.M. and M. Studdert-Kennedy (1978): "Phonetic perception", in Handbook of sensory physiology, vol. III, 'Perception', R. Held, H. Leibowitz, and H.-L. Teuber (eds.), Heidelberg: Springer-Verlag.

MacNeilage, P.F. (1970): "Motor control of serial ordering of speech, Psych.Rev., 77, 182-196.

Moll, K.L., G.N. Zimmerman, and A. Smith (1976): "The study of speech production as a human neuromotor system", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 107-127, Tokyo: University of Tokyo Press.

Turvey, M.T. (1977): "Preliminaries to a theory of action with reference to vision", in Perceiving, acting and knowing: Toward an ecological psychology, R. Shaw and J. Bransford (eds.), Hillsdale, New Jersey: Erlbaum Press.

Turvey, M.T., R. Shaw, and W. Mace (1978): "Issues in the theory of action: degrees of freedom, coordinative structures and coalitions", in Attention and performance, VII, M. Requin (ed.), Hillsdale, New Jersey: Erlbaum Press.

COMMENTS FROM THE PANELISTS

Two panelists had comments to make on the nature of the programming units. MacNeilage pointed out the potential of single motor unit research as a means for defining the nature of such units, although he also made clear that at present he and his colleagues are not attempting to posit "any straightforward relationship between these data and such concepts as the phoneme or distinctive feature". Bladon spoke somewhat more extensively on this issue. Specifically, Bladon called for the recognition of "a plurality of articulatorily relevant units", including features, phonemes and phonological syllables. He provided examples in support of each of these and then went on to say, "moreover coarticulation needs to be sensitive at times to other properties than phonologists have proposed, including a strength hierachy, including even rule-order in rapid speech forms, and including also phonetic system size (perhaps implying some sort of articulatory distance measure)". He then noted that the existence of counter-examples against all of these units might "lead into the question of perhaps whether an interesting possibility would be that different types of units might be made use of for different motor control functions".

Two panelists also took up the question of the form and function of motor control rule systems. Hirose directed his comments to these systems by pointing out that his overall aim was to "investigate the temporal organization of the speech production process", via investigations of the "relationship between the pattern of motor control signals...and the dynamic characteristics of the speech organs which act in response to the control signals". In summarizing the EMG and movement data from velopharyngeal function in Japanese presented in his paper (vol. II, p. 351-357) Hirose noted that both the EMG activity and the resultant velar movement for nasals varies predictably depending upon the class of nasal sound being produced. He states: "It can be assumed that the EMG activity for moraic /N/ is characterized by a step-like suppression and the velar movement can be regarded as a smoothed response of the second order system to it. For the initial /m/, the velar movement can be taken as a ballistic impulse response like movement. For the geminate /Nm/ there must be a positive control which can inhibit extreme lowering of the velum in spite of the longer duration of nasalization." Thus, Hirose stressed the importance of studying the relationship between EMG activity and structural movement as one method for evaluating potential motor control rules and systems. Daniloff and Tatham, on the other hand, investigated EMG activity in the production of English bilabial stops. In a reinterpretation of the original data, Daniloff reached the following conclusions, among others: First, there is "definitely an impression from the data of multiple articulatory solutions (there is no one muscle nor any one articulator that needs to move in exactly the same way from trial to trial to get a given acoustic end) and,thus, you need to know the biomechanics of an articulator in order to interpret the EMG". Secondly, and related to the first point, "coarticulation, which you expect to be extreme in a stop consonant-vowel syllable, may be optional or there may be ways to solve the coarticulation using different muscles from repetition to repetition". Finally, Daniloff stressed the close relationships which they noted between temporal characteristics of their EMG data and the resultant labial productions. Thus, in agreement with Hirose, Daniloff provided examples of the use of relationships between EMG activity and output behavior of the structures.

Two of the panelists presented views concerning the best theoretical approach to motor control research. MacNeilage stated that one of the reasons underlying his interest in single motor unit work "derived from a relative disenchantment with attempts to define the underlying abstract units of the speech production process on the basis of experimental studies of speech production". He thus wanted to provide some data about the rather high level stage of the motor unit, which he believes "defines the way the

central nervous system must encode its information", before ultimately returning to the "larger questions" of underlying units. Bladon, on the other hand, expressed concern that "the limited predictive capacity of each of these linguistic constructs (features, phonemes, and phonological syllables) have led various people to be critical". Specifically, Bladon cited both MacNeilage and Lubker in statements relevant to the lack of correspondence between production data and theoretical linguistic constructs. He suggested that "large numbers of linguistic constructs have been shown to have some relevance to the control of coarticulation and if they have come to very little effect in their operation, can you really expect all data to be supportive of any one construct?" Bladon answered his own question in the negative and expressed considerable unease at the "nihilistic" views of Moll, Zimmerman and Smith (1977) cited above in the introductory comments. In the subsequent panel discussion, MacNeilage extended his views somewhat by stating: "I think the basic state of affairs is that we have a linguistic message that we are trying to implement by a motor control system and the implementation of that message must obviously be related to the nature of that message and therefore we need to continue to struggle with the problem of what the underlying abstract forms are." And further, speaking directly to the issues raised by Bladon, he stated: "When I say that I think the theory is relatively unsuccessful, what I mean is that there is no simple set of rules that can account for the observed coarticulatory behavior. I think our problem is that we just simply have too many divergent pieces of data and we do not have a clear-cut relationship between those data and the underlying concepts like the syllable. So, we have these kinds of anomalies and we have these fairly spectacular cross-language differences in exactly how speakers handle coarticulatory events, and I would stick with my characterization that the theories have been relatively unsuccessful." In return to MacNeilages comments, Bladon agreed that there was no simple set of rules but did not think "that we should therefore conclude that a complex set of rules is a non-successful one". It would thus seem that both Bladon and MacNeilage were concerned with some form of "translation theory" approach to motor control systems in spite of some differences regarding the nature of the translation theory. Indeed, this seemed to be true in the case of

all of the present panel members. The paper by Gay and Turvey was supportive of Action Theory but since neither of them were present that view was not taken up at this point in the discussion.

Finally, Perkell provided a consideration of the relative roles of feedback and central programming mechanisms in motor control systems and in doing so pointed out that it is necessary that we "understand the way feedback works if we are ever going to come close to understanding the physiological/neurophysiological correlates of linguistic units". Perkell suggested three forms of feedback which might be important to speech motor control: (i) "oral-sensory feedback utilized over relatively long time spans in conjunction with auditory feedback to establish and maintain a subconscious knowledge of certain vocal tract states which produce sound outputs that have distinctive and relatively stable acoustic properties"; (ii) "peripheral feedback used to inform the control mechanism about changes in the frame of reference which must be taken into account in making adjustments in motor programs". Perkell discussed this second point in detail in his paper (vol. II, p. 358-364). In the present discussion he added the notion that "when a motor program is constructed and executed, it is probably accompanied by a set of expectations on the outcome of the program and feedback is likely used to compare the actual with the expected result. If a large enough mismatch is detected then adjustments have to be made in subsequent programs."; (iii) "Feedback could be used on a moment-to-moment basis in the partial control of the individual's articulatory movements or in the coordination of more or less simultaneously occuring movements of different articulators." In discussing this last form of feedback control Perkell brought in the work of Folkins and Abbs (1975) which suggests that the "peripheral reflex pathways are programmed to make on-line or moment-to-moment adjustments in commands to the articulators". He also discussed the work on head-eye coordination in monkeys which has been shown to be controlled by reflex pathways involving the vestibular apparatus. This, in turn, led him to the question: "is there anything like the vestibular apparatus for vocal tract movement coordination? In other words, in what ways might the neural organization for speech production be specialized for moment-to-moment use of peripheral feedback?" Perkell warned that in seeking answers to such questions we must

be very cautious since the experimental conditions in feedback research might cause subjects to use mechanisms which are available but not used for ordinary "ongoing overlearned speech activities". Perkell concluded by suggesting that "a great deal of movement control for ongoing adult speech production is probably accomplished through pre-programming. We use motor patterns which are stored in some kind of incomplete form and elaborated in part during pauses and in part on a moment-to-moment basis. The control mechanism could use what the motor control theorists like to call 'efferent copies' or a knowledge of ongoing motor commands which could be used to compensate for self-generated changes without having to resort to peripheral feedback. In order to account for natural variations in articulatory movement (e.g. motor equivalence) some moment-to-moment feedback function seems to be necessary. Now, this feedback function could include peripheral feedback and it probably includes feedback mechanisms contained entirely within the central nervous system (cf. the discussion by Hirose, below). The use of internal feedback in place of peripheral feedback might be part of learning how to speak and there is most likely a fluctuating use of various forms of feedback depending on the demands of the situation."

In addition to these relatively formal comments there was also some more informal discussion among the panel members, some of which has already been alluded to in the above section on theoretical approaches to questions in motor control. During this discussion Perkell pointed out that coarticulation is observed in terms of structural movement and that "we don't see the movements of features". He further observed that structural movement, using the example of the mandible, is set by goals specified as a function of time and influenced by the movement and positions of other structures such as the lips, tongue body, tongue tip and even the larynx. All of these requirements on the mandibular movement must be summed so that they "produce a set of motor goals for the mandible which is really vertical position as a function of time". Further, what seems to apply "almost universally" for such conditions is some form of "look-ahead" mechanism which checks for future goals and intervening requirements, thus allowing smooth movement from goal to goal. Perkell then notes that recent data (see discussion below by McAllister) suggests that in rounded

vowel-nonlabial consonant-rounded vowel utterances there is a trough, or reduction, in EMG activity that would not be predicted by a look-ahead mechanism. He then called for some discussion of such look-ahead mechanisms and the possibility of word or syllable boundaries to help us "nail down" such data. In response to this, Daniloff suggested that juncture which exceeds some given length of time may result in suppression of activity in certain articulators and movements towards more neutral positions. Bladon noted that although the mass of data seems in favor of articulatory spread of features such as rounding across syllable and word boundaries there may well be cases in which speakers are simply using different strategies and where boundaries "have come to be influential". However, he does feel that the weight of the evidence is to the opposite and that coarticulation does spread across such boundaries.

DISCUSSION

Since space does not permit the inclusion of all points made during the open floor discussion, only those points most relevant to the main issues raised by the panel will be taken up. Additionally, priority is given to those who were motivated enough to comply with the Congress Organizers' request to supply written summaries of their questions.

Löfqvist provided an extensive discussion of Action Theory. He pointed out that not much experimental work had yet been done within that framework but that theoretical considerations are equally important and that theoretical arguments and issues should be sorted out before starting experimental work. He said that "one of the main problems in motor control, emphasized by the Russion physiologist Bernstein, is that of reducing the number of degrees of freedom to be directly controlled". He also suggested two problems which any explanatory theory of motor control must deal with: "Movements should be made to reach a given goal irrespective of varying initial position", and "Movements should be carried out in the face of unexpected perturbations or changes in the environment." Löfqvist emphasized that both of these movement conditions must be carried out "without any lengthy search procedure". Action Theory accounts for such movement phenomena via the concept of coordinative structures, which can be "regarded as a functional grouping of muscles constrained to act as a unit.

Specified relationships between a group of muscles, expressed by equations of constraint, make the group self-regulatory." He suggested, in closing, that "the perspective of coordinative structures would lead you to predict that invariance will not be found in the individual muscles. Rather, it should be searched for in the dynamic relationships between muscles, or groups of muscles, over time.

In response to Löfqvist's comments, Lindblom asked how Action Theory accounted for the ability of the motor system to adapt to an almost infinite number of new situations while goals remain constant. Lindblom further called for the panel to clarify the term "pre-programming" which he took to mean, in general, "some kind of adaptive, creative control strategy derived on-line and involving foresight". Specifically, Lindblom called for discussion of a possible mechanism to account for such control. Hirose answered Lindblom's second question by reference to a cerebro-cerebellar loop which has been proposed by Allen and Tsukahara (1974). These authors describe a specific neurophysiologic system, the cerebro-cerebellar communication system, "the function of which is largely anticipatory, based on learning and previous experience and on preliminary, highly digested sensory information that some of the association areas receive."..."In other words, in central monitoring of efference, a copy of the motor commands sent to the muscle is monitored centrally and thus it should not wait for proprioceptive comparison." Bladon also offered some comments on Lövqvist's view of Action Theory and in doing so extended Lindblom's question concerning it. Bladon first stated that he felt that the concept of coordinative structures was quite promising. Nevertheless, he felt that there was a major problem which both Löfqvist and Lindblom had alluded to, and that was, "how do you actually investigate this, how do you test this theory, how do you compare it with what you have already?" Bladon suggested that since it has been stated that coordinative gestures involving speech are agents of coordinative structures, then perhaps experimental proof of the existence of such coordinative gestures would provide the sought after evidence. In reviewing that evidence with which he is familiar Bladon was unable to provide any direct support for such coordinative gestures and feels that the question of experimental proof for Action Theory remains an unaswered and important one.

Somewhat later in the discussion Port made a comment which was relevant to the Action Theory concept. He argued for a less limited role for timing in coarticulation theory. Specifically, he suggested that "an adequate theory of coarticulatory phenomena should probably also include explanation of examples of inherent durational effects and their compensatory adjustments as an integral part of the system--not as a different theory patched on at the end. It is even possible that by building in this kind of temporal coarticulation at the outset, we will find the entire project more tractable." Port then stated that "the notion of coordinated structure employed in action theory is intended to capture both the temporal and spatial invariants of a phonetic event. Perhaps this is a theoretical notion that could be developed to capture both the temporal aspects of the spatial position of articulators as well as the inherent temporal structure of segments and prosodies."

Turning in another direction, McAllister responded to Perkell's question (see above) concerning the failure of "look-ahead" models to account for the observed "trough" in recently reported EMG data. McAllister showed simultaneous movement and EMG data from labial function during the production of rounded vowel--nonlabial consonant string--rounded vowel utterances. The nonlabial consonant strings consisted of one, four and six consonants. These data clearly showed troughs, or relaxations, in both the EMG activity and in the lip rounding, the most interesting point being that the relaxations occured at the boundary between the offset of the consonant string and the onset of the second vowel. McAllister agreed with Perkell that such data are incompatible with previous descriptions of the look-ahead mechanism, and stated that he is particularly "hard pressed to explain the location of the trough." He suggested that there may be "a critical acoustic boundary" at that point which demands a "neutralization" of rounding.

Ohala suggested that our search for underlying units would perhaps be facilitated by examining cases where coarticulatory behaviors were "clear" rather than "smeared". Specifically, he presented a number of examples of cases, in Swedish and in English, where coarticulatory behavior was time-locked to phonemes.

As a final point in this summary of the discussion from the floor, the comments made by Porter may be appropriate. Porter called for considering production and perception phenomena more closely together rather than as distinct fields of study. He felt that this would aid us in "terms of understanding perception and also in understanding the role of feedback in the control of output". Porter extended his argument via Action Theory by noting that somewhere between "abstract phonetic entities and the more concrete properties of motion and acoustics" there must be an "interface and a common code". That is, a common code to the exclusion of a translation theory. A code that functions both in production and in perception.

Very little summary is required for the above comments. It seems very clear that answers are being sought and that there is a healthy amount of controversy. The seeking and the controversy suggest that researchers in the field of motor control are, indeed, working toward those goals stated by Lindblom in his plenary lecture: that "phoneticians should extend their inquiry into the sounds of human speech to ever deeper physiological and psychological levels using speech as a window to the brain and mind of the learner, talker and listener", and, further, that we should expect "more complete, theoretical models and computer simulations to be proposed for speech production, speech understanding and speech development that match the present quantitative theory of speech acoustics in rigor and explanatory adequacy".

References

Allen, G.I. and N. Tsukuhara (1974): "Cerebrocerebellar communication system", Physiol.Rev. 54, 956-1006.

Folkins, J. and J. Abbs (1975): "Lip and jaw motor control during speech", JSHR 19, 207-220.

Moll, K.L., G.N. Zimmerman, and A. smith (1977): "The study of speech production as a human neuromotor system", in Dynamic aspects of speech production, M. Sawashima and F.S. Cooper (eds.), 107-127, Tokyo: University of Tokyo Press.

SYMPOSIUM NO. 7: THE RELATION BETWEEN SENTENCE PROSODY AND WORD PROSODY

(see vol. II, p. 375-430)

Moderator: Eva Gårding

Panelists: Arthur S. Abramson, Gösta Bruce, Johan 't Hart, Eunice V. Pike, Nina Thorsen, and Kay Williamson

Chairperson: George D. Allen

EVA GÅRDING's INTRODUCTION

The purpose of the symposium is to discuss the relation between sentence prosody and word prosody in different prosodic systems, with the aim of tracking down universal features and tendencies in this relation. A more general goal is to contribute to a common framework for the description of prosodic phenomena. Since one of the symposia deals with length, such features have not been included here. To secure a broad treatment of the topic, a number of specialists of various prosodic systems were invited to be members of the panel. They represent Thai (Abramson), Amerindian languages (Pike), Nigerian languages (Williamson), Swedish (Bruce), Danish (Thorsen), Dutch ('t Hart), and Czech (Jánota).[1]

In volume II p.375 I proposed a terminology and suggested some points for discussion. I shall first elaborate on these points (1.1 - 1.4). Next follow summaries of the panelists' comments to their written contributions (2) and then an account of the discussion, ordered by subject (3.0 - 3.3). With this order some of the contributions have had to be split up under different headings. Finally I try to give a short evaluation of the symposium (4).

1.1 Basic units[2]

The first basic concept which is fundamental to our discussion is sentence intonation. Everybody on the panel agrees that an observed pitch pattern is equal to sentence intonation plus word intonation. But there are different views about what these two components really are and how they should be extracted from an observed curve. For those who treat tone languages and 2-accent languages, sentence intonation seems to be a broad general fea-

---

1) Přemysl Jánota was unable to attend the congress.

2) See footnote on page 293.

ture (called global in what follows), possibly combined with a local feature. These features express the illocutionary character of an utterance, for instance, statement or question. They can be manifested as downdrift or absence of downdrift with or without some consistent local glide. The ups and downs determined by the tones and accents are imposed on this pattern.

For 't Hart and Collier in their analysis of Dutch, however, intonation is the total intonation pattern including the rises and falls over the accents. Word prosody is lexical accentuation and it only determines the timing of some salient parts in the pattern. Palmer (1922), Bolinger (1958), and O'Connor and Arnold (1961) have described the intonation of English in a similar way.

It seems clear that the existence of these two radically different interpretations does not facilitate our task.

In connection with the concept sentence intonation we should perhaps ask ourselves the following questions:

Are the prosodic systems really so different that they have to be analysed differently?

Is a compromise possible so that sentence intonation can be given the same meaning in different prosodic systems?

Are there any languages for which the decomposition into word prosody and sentence prosody is meaningless?

Is there perhaps a need for a smaller unit between sentence and word, such as phrase?

The second concept important for our discussion is sentence accent. Even here there is fundamental disagreement. About half of the panel take sentence accent to be an accent feature expressing the focus of a sentence which can signal semantically or emotionally important words. In widely different prosodic systems, sentence accent has been reported to have similar manifestations: increased duration and amplitude in combination with a special pitch pattern. Most often sentence accent occurs on the accented syllable of the word in focus but it can also have a separate manifestation on a later syllable. Such cases have been reported by Eunice Pike for Ayutla Mixtec and Acatlan Mixtec (p.414) and by Gösta Bruce and myself for Swedish dialects (p.388). As a rule the tone languages listed by Eunice Pike have sentence accent. Kay Williamson, on the other hand, does not need the concept for her description of Nigerian tone languages and Nina Thorsen as-

cribes the prominent accents elicited from Copenhagen speakers to emphasis or contrast.

't Hart and Collier do not separate a special sentence accent from other accents. All pitch movements in combination with accented syllables are sentence accents. This is consistent with their view of intonation.

The sentence accent has been very useful in the analysis of Swedish intonation and I am ethnocentric enough to think that it should be useful generally. I therefore suggest that we discuss the relevance and usefulness of sentence accent. Also here we might need an intermediate level between word and sentence. A parallel term to phrase intonation would be phrase accent.

The other basic units are of course accents and tones but competing descriptions of tones and accents, although abundant in the literature,[1] are not to be found in the contributions to this symposium. They may come up in the open discussion, however.

1.2 Extraction of the phonetic correlates of basic units

Suppose now that we have some idea of the linguistic nature of the basic prosodic units at sentence and word level. How should we extract their phonetic correlates from observed pitch patterns? To do this extraction it seems necessary to consider utterances in which sentence prosody and word prosody are varied in a systematic fashion. This is the method which has been used by Gösta Bruce. The method may lead to basic forms that are not always directly observable in a given pattern. For Swedish dialects we have in this way extracted four different manifestations of sentence accent which are extremely useful in generating and explaining the different types of intonation in Swedish dialects.

For Abramson it is the citation form which contains the phonetic correlates of the basic tone and this form is then perturbed by sentence prosody and adjacent tones.

There are hardly any competing views about the phonetic correlates of tones but for accents the pendulum has swung between pitch and intensity. For a long time now it has been customary to regard all accents as pitch accents. I found it very refreshing to see the data presented by Fujisaki and his collaborators in a poster session at this congress (Fujisaki et al., 1979a). The data seemed to reestablish some of the importance of intensity for English accents as compared to Japanese ones.

---

1) See e.g. references in Leben (1978).

For sentence intonation, various auxiliary lines have been proposed. 't Hart and his collaborators have used a baseline joining local minima in a curve, only for them it does not represent sentence intonation.[1] Nina Thorsen joins points (lows) representing stressed syllables. For Swedish we have used a more complex construction of baselines and toplines (Bruce and Gårding, 1979). Common to all these constructions is a baseline whose steepness is determined by the length of the phrase. In Fujisaki's intonation model, which he showed during the discussion ensuing the report on perception, the baseline is independent of the length of the utterance (Fujisaki et al., 1979b). I have asked him to give a brief demonstration of the pertinent parts of his intonation model at the end of the time allotted to the panelists.

To sum up my questions under this point (1.2):
I suggest that we discuss various methods for the extraction of the phonetic correlates of the prosodic units.
How should this extraction be done and to what purpose?
Are principally different methods possible?
And what are the phonetic correlates of the basic units, sentence intonation, sentence accent, lexical tone, lexical accent?

1.3 Interaction between sentence prosody and word prosody

Let us now assume that we have extracted the phonetic correlates of the basic units of sentence prosody and word prosody. To generate perceptually correct pitch patterns we must know how these units interact. And here finally we come to the main theme.

Generally speaking, sentence prosody precedes and sets the scale for word prosody. This must be a true universal. For instance, downdrift influences everything on its way, and in Swedish, sentence accent influences all preceding and following word accents.

Apart from the interaction between sentence prosody and word prosody there is also interaction between adjacent units in the utterance, usually called tonal coarticulation and described by tone rules (Hyman and Schuh, 1974; Schuh, 1978).

I suggest the following points of discussion under 3:
Is the order sentence prosody, word prosody a true hierarchy?
And at the sentence level, is sentence intonation primary to sentence accent?
Are there any general principles governing tonal and accentual coarticulation?

---

1) 't Hart modifies this statement: The baseline is not the only manifestation of sentence intonation.

1.4 Additional questions

Here I collect questions which are marginal to the main theme. How does one determine if the basic prosodic unit for a word is a tone or an accent? According to Eunice Pike it is possible to determine if a given High represents an accent or a tone by studying its effect on vowel quality. Accented syllables have full vowels and unaccented vowels are reduced. Also accented consonants are affected. High tone, on the other hand, has no influence on vowel quality.

Accent also affects duration in a drastic way. In Swedish an accented syllable is more than twice as long as an unaccented one, whereas tone only has a marginal effect on duration.

According to many linguists, e.g. Larry Hyman (1975, p. 207 ff.) the difference between tone and accent is a linguistic one, not a phonetic one. I think that this point should be debated further. Tone and accent seem to have quite different contextual effects, difficult to explain without some difference of physiology.

2. COMMENTS FROM THE PANELISTS

Arthur Abramson emphasizes that the five tones of Thai are essentially preserved in connected speech.[1] He goes on to give an example which shows that the declination over an utterance is 30% of a woman's voice range, with the topline responsible for a larger amount of the declination than the baseline. Sentence accent is perhaps not as adequate a notion for the description of Thai as syntactic groupings in which phrase breaks are signalled by prosodic variation.

Eunice Pike summarizes ways in which pitch is used in the languages she has studied. It signals contrasts between lexical items, segments a stream of speech into words and clauses, marks sentence stress and conveys attitudinal meaning. Eunice Pike exemplifies these functions in various languages. In Marinahua of Peru a high tone will be still higher and a low tone lower under sentence stress. In Mikasuki of Florida tones are modified downward to mark boundaries between words and upward to mark bound-

---

1) According to Gsell (1979) the distinctiveness of tone in Thai is very much reduced in connected speech. There are only certain positions, comparable to accented syllables, in which the tones retain their distinctive power. - This publication contains a lot of other information relevant to the theme of this symposium.

aries between phrases. In Eastern Popoloc of Mexico a final upglide marks politeness as opposed to the unmarked neutral ending with a glottal stop. (For references see Vol. II p. 416). In Fasu high tone and low tone contrast lexical items only in stressed syllables, the unstressed syllables carry attitude or sentence intonation. A special voice quality is used in talk with spirits.

Kay Williamson calls attention to tonal modifications due to grammatical constructions which in her present view were underemphasized in her earlier contribution (p. 424). With fewer minimal pairs there is more freedom for extensive variation without causing ambiguity. One of the languages has some dialects which could be called pitch accent systems. Such a system may have developed as follows. Series of high tones have gone low and the surviving highs have become - phrase accents! Kay Williamson exemplifies global and local effects in connection with sentence type. Global manifestations are downdrift, a cancelling of downdrift or a raising of highs so as to increase intervals. One example of a local effect is that in Igbo the normal pronominal repetition of a subject at the beginning of a phrase has a high tone in the statement and a low tone in the question. In all other cases the local effect occurs at the end of the sentence with an opposition between statement and question. There is a final high for statement as opposed to low for question in some of the languages, which goes to show that the connection of high with question and low with statement is not a universal one.

Gösta Bruce shows a Stockholm Swedish pitch contour with six word accents surrounding a sentence accent in the middle of the utterance (Fig. 1). This figure shows that there are two contextual variants of one and the same accent, depending on their position relative to the sentence accent, rise-falls before the sentence accent and mere falls after it. Statement intonation is represented by the downdrift. The extent of this downdrift for a given speaker

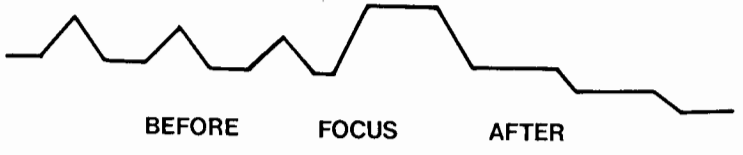


Fig. 1. Downdrift in Swedish. Stylized pitch contour of a Swedish utterance. From Gösta Bruce. Work in progress.

seems to be independent of the length of the utterance. However, the figure, assumed to be typical in this respect, shows that the actual course of the downdrift pattern has a very gentle slope before the sentence accent and a steeper, terrace-shaped downdrift afterwards. The figure sums up some important aspects of the interaction between sentence prosody and word prosody. Sentence intonation sets the scale for accentuation and accentuation determines the time course, in this case of the downdrift.

Nina Thorsen needs two prosodic units between word and sentence, the stress group, defined as the stressed syllable and the succession of unstressed ones, and a prosodic phrase group consisting of several stress groups. In her prosodic system there are two components which do not interact. Stress-group patterns are simply superimposed on the intonation contour which in her model is described as a line joining the stressed syllables. Nina Thorsen further discusses problems of definition when she applies this view to utterances with emphasis for contrast. She prefers to think that with emphasis the utterance is reduced tonally to a one-stress utterance. With this interpretation the difference between statement and question lies mainly in the stressed syllable and the post-tonic syllables.

Johan 't Hart underlines that in his and his collaborators' analysis of Ducth, declination is part of the intonation but not the only manifestation of it. Word prosody is lexical accentuation and sentence accentuation is represented by the pitch accents in the sentence. Sentence intonation has a higher place in the hierarchy. Reference to the communicative function has been avoided. Intonation patterns are not connected with linguistic categories such as statements, questions, wishes or commands, but represent classes of melodical shapes distinguished by the listener.

Hiroya Fujisaki in an extra contribution invited by the moderator, describes a model for Japanese intonation. It is, he says, principally similar to an intonation model proposed by Öhman (1967). In logarithmic scale all F_o patterns are sums of two components, a baseline component (called voicing component) corresponding to sentence prosody and an accent component. Fujisaki showed a figure (Fig. 2) that strengthens his view that the time constant of the baseline is not affected by sentence length. In longer sentences the speaker resets his baseline at one of the major syntactic boundaries. A general observation is that with an

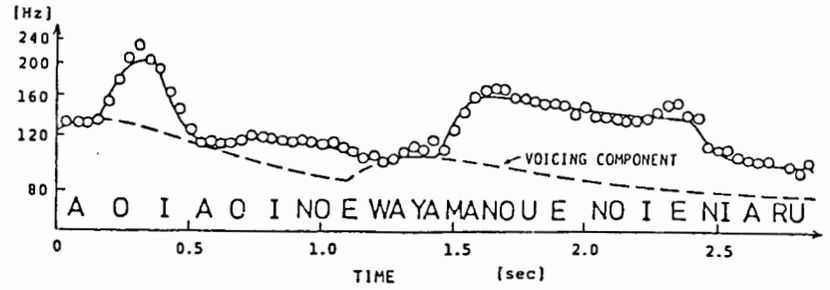


Fig. 2. Analysis by synthesis of a Japanese F_o-contour with two voicing (baseline) commands. From Fujisaki et al. (1979b).

absolute scale the height of accentual F_o peaks over the baseline decreases towards the end of a declarative intonation contour. In logarithmic scale, however, the peaks have approximately the same height over the baseline. This analysis can lead to a simpler and more illuminating interpretation of prosody.

3. DISCUSSION

In this section I have chosen to organize the discussion by subject. Consequently one intervention may occur in several places. I have followed the terminology of each discussant, inserting my earlier suggested term within parentheses. Terminological remarks, in particular those with a bearing on typology, have been collected under point 3.0. Since all the additional questions (1.4) concern the basic units and their correlates, they have been referred to 3.1 and 3.2. Otherwise the points for discussion follow the suggested outline. The discussion typically begins with the panel, proceeds with the respondents from the audience and ends with the panelists' responses.

3.0 Terminology

Irmgard Mahnken wants the terminology to show the non-isomorphic character between grammatical and prosodic units.

William Moulton offers a list of terms useful for the description of different prosodic systems. Three uses of pitch and stress, lexical, morphological and syntactical, can be combined in different ways. William Moulton also underlines the need to distinguish between gradient versus discrete pitch and stress signals.

3.1 Basic units

All the panelists agree on the usefulness of an intermediate unit between sentence and word level.

For the description of a Subject Object Verb language, Kay Williamson uses the concept tone group. This tone group is syntactically determined. Within such a group the first word sets the pattern for the whole group. For the group Object Verb the verb loses its own pattern and follows that of the object. In the dialects mentioned earlier, where only one High per group survives, normally the last one, group accent might be an appropriate term.

Also Johan 't Hart advocates the idea of introducing groups into the descriptive framework.

Eva Gårding argues that in the data presented by Arthur Abramson for Thai (p.383) one can find phrase accents manifested as increased amplitude and length and in the same utterance also something that looks like a sentence accent with an even more prominent increase of length and amplitude. In her own dialect of Swedish there are similar phenomena. Lexical restrictions on the pitch pattern in an accent language like Swedish make it perhaps more convenient to signal a syntactic unit by a phrase accent, expressed by increased amplitude and length rather than by a particular pitch configuration, as for instance in the Dutch hat pattern.

Arthur Abramson agrees with this interpretation of phrase accent in his material but he is not happy with the notion of sentence accent, which is determined by the whole discourse.

René Gsell gives a linguistic functional definition of tone, accent and sentence which he missed in the panelists' discussion. (This critique was repeated by other discussants, e.g. Mahnken, Moulton and Carton.) Tone is a paradigmatic mark of morphemes and words. Accent is a syntagmatic mark and the function of accent is the grouping of morphemes into words and at a higher level, of words into tagmemes and larger phrase constituents. In the symposium sentence accent has been used for emphasis and focus, which are two different things. From a linguistic point of view sentence accent is mainly phrase accent, the culminative mark of a higher constituent. Intonation is a still higher level of integration by which tagmemes or constituents are grouped into sentences.

Vichin Panupong demonstrates how in Thai sentence intonation can be signalled by final tone-bearing particles. One such particle is ka which modifies the total meaning of a sentence from e.g. statement to question by means of one of four possible tones. Sentence intonation can be carried by a final word as well. Final particles are also used to mark boundaries.

Sieb Nooteboom comments on the confusion between pitch accent in the Dutch analysis as compared to sentence accent in the Swedish one. The Swedish picture of one accent determined by focus surrounded by a number of smaller ripples caused by other accents (Fig.1) may correspond to just one pitch accent in Dutch determined by focus without any pitch manifestation of the other accents. Gösta Bruce has analysed sentences with only one semantically determined pitch accent whereas 't Hart (p.398) shows sentences with a number of semantically determined pitch accents. The question is what would happen in Swedish in a comparable situation, i.e. in a sentence with several semantically determined pitch accents.

Fernand Carton points out that even within one language there are problems of description. He needs the notion of accent (as do other analysts) for his study of dialects in the north of France where accent is still contrastive. Other analysts, as e.g. Mario Rossi, claim that there is no accent in modern French since it has only demarcative (syntactic) function. A common theoretical framework is needed, which takes functional aspects as well as the existence of different factors into account. A constant check on the interplay between form and substance is needed at all stages of the analysis and perceptual tests are crucial.

Alan Cruttenden is disturbed by the continued use of such simple categories as statements and questions for sentence intonation.

Barbara Prohovnik thinks that an intermediate unit like prosodic phrase might have a bearing on the definition of the word and the sentence.

Lisa Selkirk with experience from comparative work in French and English wants to posit an intermediate level which has a syntactic definition.

Philippe Martin wonders how phoneticians can say that there are well formed sequences of pitch accents, as for instance in Dutch, if they reject any relation between syntax and sentence intonation.

Responses to 3.0 and 3.1

Gösta Bruce answers Sieb Nooteboom that there may be two or three sentence accents in the same Swedish utterance.

Eva Gårding is of the opinion that all panelists agree with René Gsell on the importance of function in a linguistic analysis.[1]

Kay Williamson in response to William Moulton's typological suggestions says that at least nine combinations of pitch and stress are needed. We speak of tone languages, stress languages and pitch accent languages, but we need more categories for the languages described in Eunice Pike's contribution, where both stress and pitch are contrastive. There are in addition at least two types of tone languages, the syllable-tone type and the word-tone type. To sum up, we need a rather more complex typology than the ones suggested earlier.

Eva Gårding reassures Alan Cruttenden that the members of the panel are well aware of the existence of a variety of sentence intonation types. The reason there is so much talk of statement and question intonation in the contributions is that the purpose of the symposium is to study the relation between word and sentence prosody and that this can be done safely in the statement and question types since they are well established in prosodic systems and easily elicited from speakers.

3.2 Extraction of the phonetic correlates of basic units

3.2.1. Citation forms versus other forms

According to Gösta Bruce citation forms would be insufficient for a thorough analysis of an accent language like Swedish. A Swedish citation form is a very complex pattern containing contributions from several linguistic variables, word accent, sentence accent, sentence intonation and terminal juncture. His results have been obtained by comparing words in different prosodic contexts. In this way it has been possible to decompose the classical double-peaked Accent 2 pattern of e.g. Stockholm Swedish into a word accent fall, a sentence accent rise and a terminal juncture fall.

Arthur Abramson defends the use of citation forms, partly for practical reasons - they are easy to elicit and measure - and partly for psychological reasons - children tend to learn one-word

---

1) I was too rash here. Gsell and Moulton and others requested a functional definition of the concepts under discussion. It should have been said from the beginning that the basic units were intended to be useful and efficient in the analysis and synthesis of prosody. In this capacity they are not necessarily functional units in the classical sense.

utterances and hence citation forms.

Alan Cruttenden gives an example from one variety of Panjabi which supports the view that the basic form of pitch accent should be derived from connected speech rather than citation forms. In connected speech a two-way pitch accent distinction involves a clear deviation downwards or upwards respectively in a particular intonation pattern, whereas in citation forms the distinction is very complex.

Eunice Pike finds it very important to remember in an aural linguistic analysis that lexical tones may be modified by sentence intonation or sentence stress. One trick in such an analysis is to ask for three items and have the words you want to contrast as number one and two. These two will then have a chance to have the same intonation pattern whereas the last item will have terminal intonation. To separate sentence accent from lexical tone it is advisable to have at least two words in a sequence. One of these words will then have the sentence accent and the other words will carry only tone.

3.2.2. Methods for the extraction of basic forms

At least four methods have been mentioned in the contributions, elicitation of citation forms (Abramson), comparison of prosodic variables in different contexts (Bruce, Pike), analysis by perception ('t Hart), and analysis by synthesis (Fujisaki).

Edward Purcell makes a request for more statistically based approaches to modelling tone and intonation, by using e.g. polynomial regression. It might then be possible to solve equivalence problems like the Dutch and Swedish sentence accent.

Yukihiro Nishinuma points out that an intonation model has to take the integration of independent acoustic parameters into account as well as the effect of masking at different levels.

Responses to 3.2.2

Johan 't Hart argues that the most important need is not statistics but a large inventory of intonational possibilities and perceptual testing. He would like to know if Hiroya Fujisaki is as concerned about the fit between synthetic and perceptual patterns as he is about the fit between synthetic and acoustic ones. As for logarithmic versus linear scale he does not think it matters much in short utterances.

Arthur Abramson is in sympathy with the use of polynomial regression but finds it most often sufficient to form hypotheses based on the acoustic manifestations and to test these hypotheses perceptually.

3.2.3 Phonetic correlates

a) Sentence intonation and downdrift

Nina Thorsen points out that in her Danish material downdrift is evenly distributed over the utterance. The downdrift does not occur only in connection with the accented syllables as shown in Bruce's figure (Fig. 1). Also, the range varies with the length of a sentence within certain limits. Contrary to Fujisaki's model for Japanese, the downdrift in her material is a linear function of the length of a short utterance. In long ones there is a resetting of intonation in connection with syntactic boundaries. She referred to the figure (Vol. II p. 417) where it appears that the height of the post-tonic syllables above the "baseline" does decrease toward the end, even with a logarithmic scale.

Gösta Bruce ascribes the difference between the distribution of downdrift in Swedish and Danish to the different use of sentence accent. In standard Swedish a normal neutral utterance will have sentence accent on the last accented word whereas in Danish and perhaps also in Southern Swedish dialects there is no obligatory rule. The range of the downdrift has appeared to be constant in sentences with two, three and four accented syllables.

Osamu Fujimura mentions work on pitch synthesis conducted by Janet Pierrehumbert at Bell Laboratories. It is somewhat similar to the work reported by Hiroya Fujisaki. The algorithm is based on specifications of pitch peaks representing relative prominence with options for low-tone stress. Nuclear tones fall below the baseline and postnuclear tones are neutralized. Pitch declination is a descending time function with resetting at major phrase boundaries (see Pierrehumbert, 1979).

Hiroya Fujisaki agrees with Johan 't Hart that the scale is not so important within a small range but for longer sentences the distinction is very clear. In answer to Nina Thorsen he says that there may be many language-specific points in prosody. He strongly agrees with Edward Purcell about the need for analytic and quantitative methods in the analysis of the production and perception of prosodic phenomena.

b) Accent versus tone and accent versus stress

In her description of the dialects of Ịzọn Kay Williamson tries to show that there is a gliding scale between tone-dialects

and accent-dialects with a very narrow cross-over zone. In general, [and this is consistent with Eunice Pike's description, EG] the more you have a tone language, the more things are symmetrical, and the more you have an accent language, the less things are symmetrical. The accents have more prominence and other things get reduced in relation to it. Perhaps this is the reason why it is easier to talk about sentence accent in accent languages than in tone languages.

René Gsell says that from a functional point of view the Scandinavian languages, even Danish, are tone languages. The 'stød' acts as an intonation depressor and is a clear example of interaction between word and sentence prosody.

Yukihiro Nishinuma (and also Irmgard Mahnken) find that in the discussion of intonation too much emphasis is put on F_0, although everybody who has worked on automatic intonation detection knows that F_0 is not sufficient.

Ivan Fónagy presents the acoustic correlates of a Hungarian phrase akar, a kar (with accent on the first and second syllable respectively) as a statement and as a question in normally intoned and whispered speech, by which he wants to show that pitch accent is not an appropriate term for the acoustic correlates of the accent. As a term he prefers stress.

Responses to 3.2.3

Eva Gårding agrees with the view that too much emphasis has been put on F_0. This trend seems to have been weakened lately.

Arthur Abramson points out that apart from fundamental frequency and amplitude variations there are also other cues that may have signal value, creaky voice and various other forms of laryngeal constriction.

3.3 Interaction between sentence prosody and word prosody

3.3.1 Hierarchy

Three views are represented at the symposium: Sentence prosody is primary (e.g. Bruce, 't Hart), lexical prosody is primary (Abramson), and sentence prosody and lexical prosody are at the same level. The last view is implied by the model presented by Hiroya Fujisaki. Here the word-prosodic part and the sentence-prosodic part are extracted simultaneously from an observed curve and may therefore be regarded as belonging to the same level of the hierarchy. The final F_0 contour is the sum of these two parts.

Arthur Abramson's feeling is that lexical prosody must be paramount in a tone language. In the mental lexicon the storage form must carry the tone as part of the morpheme. When these tones are strung together in connected speech a particular intonation is superimposed.

According to Johan 't Hart there is a higher hierarchical position for sentence intonation.

René Gsell claims that with the definitions he has given earlier (see 3.2) it is easier to understand interaction. At each level a higher constituent mark modifies lower constituent marks. Intonation dominates sentence accents, sentence accents dominate the word accents and so on. The phonetic characteristics of lower marks are not indifferent to the grouping of higher layers.

Einar Haugen remarks that the Scandinavian word accents are part of the stress pattern of the sentence and always to be seen in relation to the whole utterance. Therefore, to ask whether the word or the utterance is primary is a chicken-and-egg kind of question. You cannot say any Swedish or Norwegian word without having both tone and sentence intonation. They are stored with the word. Every native knows which tone a word has, although it never occurs without sentence intonation. Accent 2 has to be interpreted as a perturbation of the unmarked sentence intonation.

Responses to 3.3.1

Eva Gårding refers the conflicting views about the hierarchical relation between sentence prosody and word prosody to different points of departure. To work out a program for pitch synthesis by rule you need a rough idea of the sentence intonation, i.e., where to start on the frequency scale etc. So with this aim in view it is very natural to regard sentence intonation as primary. But with a psycholinguistic approach you are interested in the forms stored in the memory and the citation forms become primary in your hierarchy. These will then be perturbed by sentence prosody at some secondary level, the phrase or the sentence level.[1]

3.3.2 Contextual interaction

Arthur Abramson points out that sandhi phenomena are phonological and have nothing to do with the interaction treated in this section.

---

1) Gabrielle Konopczynski suggests in a written contribution submitted after the symposium that one should look for a hierarchy by studying in detail how children acquire tone languages.

Gösta Bruce's figure (Fig. 1) gives a good example of interaction between sentence accent and word accents on the one hand and sentence accent and sentence intonation on the other.

George Allen is interested in the deletion of postnuclear accented syllables in an English phrase. This pattern seems to be acquired quite early by children, at the age of 30 to 36 months.

Osamu Fujimura remarks that problems of accentual patterns, such as interaction between sentence accent and lexical accents have been discussed extensively in the traditional linguistic literature in Japanese. He wants to call attention to McCawley's (1968) monograph.

Perceptual tests have shown that the pitch declination effect is compensated for by listeners when they judge the height of accent peaks (Pierrehumbert, 1979).

Ana Tataru exemplifies different relations between word accent and sentence accent in Romanian on the one hand and English and German on the other. Such differences are of great pedagogical interest.[1]

3.3.3 Word prosody restricting sentence prosody

Gösta Bruce comments on the often heard assumption that a speaker of an accent language like Swedish is less free in his or her use of pitch as an expression of sentence type and attitude than a speaker of another language, like Dutch for instance. There are restrictions in the possible use of pitch movements locally but globally you are free to express other aspects of intonation.

Johan 't Hart points out that in Dutch there are also restrictions. After a rise, pitch has to come down again to be ready for the next rise. He refers to the examples given in his contribution (p.398) to show that there are also restrictions in the placement of the pitch movement which may to some extent be determined by the syntactic boundaries.

Einar Haugen reminds the audience of Otto Jespersen, who claimed that Norwegians and Swedes were unable to express nuances of feeling as well as Danes, because of the tones. It was to disprove this point that Einar Haugen went into the study of tone!

---

1) Paul Schäfersküpper in a written contribution points out that in German, sentence accent operates over larger domains than the syllable.

4. MODERATOR's AFTERTHOUGHTS

The aim of the symposium was to discuss word prosody and sentence prosody and the relation between them. Although precise results or general agreement were not to be expected, the symposium has contributed new material and well-taken points, and it has put some important questions into focus. I shall list some of them here.

It seems that even a large number of prosodic systems, as varied as those represented at the symposium, are sufficiently similar to be treated in a common framework, and that the dichotomy between word prosody, which I would now prefer to call lexical prosody, and sentence prosody, including phrase prosody, is useful even in languages whose lexical prosody is predictable from simple rules.

To find the basic units of the dichotomy we need data from all levels of analysis on which models can be based. I especially want to stress the need for simple but strict generative models. These models should aim at simulating observed patterns of pitch (F_o), intensity and duration. Without such models the interaction between word prosody and sentence prosody cannot be stated with a sufficient degree of precision.

The symposium has given strong evidence for some general tendencies in the interaction between sentence prosody and word prosody. Declination or downdrift has been observed for many languages representing a variety of prosodic systems. We have seen in the Swedish material how this gradual downdrift may be checked by an intervening sentence accent (Fig.1). It is quite possible that there are phonological systems where downdrift is masked by a late obligatory sentence or phrase accent.

Accent reduction brings out an interesting tendency. After the sentence accent (nuclear stress) all following accents tend to be reduced. There is evidence for this from Danish, Dutch, Swedish and Japanese (see Fujimura's intervention). This may be one of the asymmetries that Kay Williamson and Eunice Pike found typical of an accent language as compared to a tone language. A worthwhile project would be to explore the physiological background of this effect.

It has often been observed that the heights of equally strong accents decrease over a declining baseline. As pointed out by

Hiroya Fujisaki, however, their absolute heights are proportional to that of the baseline. This may be a universal.

Are there any general principles behind tonal and accentual coarticulation? This question was left unanswered. One of the reasons may be that these relations can only be studied together with durational aspects which were not included in the topics of the symposium.

References

Bolinger, D.L. (1958): "A theory of pitch accent in English", Word 14, 109-149.

Bruce, G. and E. Gårding (1978): "A prosodic typology for Swedish dialects", in Nordic prosody, E. Gårding, G. Bruce, and R. Bannert (eds.), 219-228, Travaux de l'Institut de linguistique de Lund 13.

Fujisaki, H., K. Hirose, and M. Sugitō (1979a): "Comparison of word accent features in English and Japanese", Proc.Phon. 9, 376, Copenhagen: Institute of Phonetics.

Fujisaki, H., K. Hirose, and K. Ohta (1979b): "Acoustic features of the fundamental frequency contours of declarative sentences in Japanese", RILP 13, 163-173.

Gsell, R. (1979): Sur la prosodie du Thai standard: Tons et accent, Paris: Université de la Sorbonne Nouvelle.

Hyman, L. (1975): Phonology: theory and analysis, New York: Holt, Rinehart and Winston.

Hyman, L. and R.G. Schuh (1974): "Universals of tone rules: Evidence from West Africa", Linguistic Inquiry 5, 81-115.

Leben, W.R. (1978): "The representaiton of tone", in Tone, V. Fromkin (ed.), 177-219, New York: Academic Press.

McCawley, J.D. (1968): The phonological component of a grammar of Japanese, The Hague: Mouton.

O'Connor, J.D. and G.F. Arnold (1961): Intonation of colloquial English, London: Longmans.

Öhman, S.E.G. (1967): "Word and sentence intonation: A quantitative model", STL-QPSR 2-3, 20-54.

Palmer, H.E. (1922): English intonation, Cambridge: Heffer.

Pierrehumbert, J. (1979): "The perception of fundamental frequency declination", JASA 66, 363-369.

Schuh, R.G. (1978): "Tone rules", in Tone, V. Fromkin (ed.), 221-256, New York: Academic Press.

SYMPOSIUM NO. 8: THE PERCEPTION OF SPEECH VERSUS NONSPEECH

(see vol. II, p. 431-489)

Moderator: David B. Pisoni[1]

Panelists: Anthony E. Ades, Pierre L. Divenyi, Michael F. Dorman, Dominic W. Massaro, and Quentin Summerfield

Chairperson: Arthur S. Abramson

DAVID B. PISONI's INTRODUCTION

Historically, the study of speech perception may be said to differ in a number of ways from the study of other aspects of auditory perception. First, the signals used to study the functioning of the auditory system were simple and discrete, typically varying along only a single physical dimension. By contrast, speech signals display very complex spectral and temporal relations. Although speech signals have also been varied along single physical dimensions, the perceptual consequences of such manipulation have not always followed from "equivalent" stimulations of a nonspeech nature. Alternatively, we may presume that the complexity of the spectral and temporal structure of speech and its variation is one additional source of perceptual differences between speech and nonspeech signals. Second, most of the research dealing with auditory psychophysics over the last thirty years has been concerned with the discriminative capacities of the sensory transducer and the functioning of the peripheral auditory mechanism. In the case of speech perception, however, the relevant mechanisms are assumed to be centrally located and intimately related to the more general cognitive processes that involve the encoding, storage and retrieval of information in memory. Moreover, experiments in auditory psychophysics have typically focused on experimental tasks and paradigms that involve discrimination rather than identification or recognition, processes thought to be most relevant to speech perception. All in all, it is generally believed that a good deal of what has been learned from research in auditory psychophysics and general auditory perception is only marginally relevant to the

---

1) David Pisoni could not be present at the congress and Michael Studdert-Kennedy acted as moderator at the meeting. David Pisoni is author of the introduction below.

study of speech perception and to an understanding of the underlying perceptual mechanisms. This situation has changed for the better in recent years as shown by the work of Dr. Divenyi and other psychophysicists who have become concerned with questions of speech perception. Despite these obvious differences, investigators have been interested in the differences in perception between speech and nonspeech signals. That such additional differences might exist was first suggested by the report of the earliest findings of categorical discrimination of speech by Liberman and his colleagues (1957). And it was with this general goal in mind that the first so-called "nonspeech control" experiment was carried out by Liberman and his colleagues (1961) in order to determine the basis for the apparent distinctiveness of speech sounds. In this study the spectrographic patterns for the /do/ and /to/ continuum were inverted producing a set of nonspeech patterns that differed in the onset time of the individual components. The results of perceptual tests showed peaks in discrimination for the speech stimuli replicating earlier findings. However, there was no evidence of comparable discrimination peaks for the nonspeech stimuli, a result that was interpreted at the time as further evidence for the distinctiveness of speech sounds and the effects of learning on speech perception. Numerous speech-nonspeech comparisons have been carried out over the years since these early studies, including several of the contributions to the present symposium. For the most part, these experiments have revealed results quite similar to the original findings of Liberman et al. Until quite recently, research reports have confirmed that performance with nonspeech control signals failed to show the same discrimination functions that were observed with the parallel set of speech signals (Cutting and Rosner, 1974; Miller et al., 1976; Pisoni, 1977). Subjects typically responded to the nonspeech signals at levels approximating chance performance. In more recent years, such differences in perception have been assumed to reflect two basically different modes of perception--a "speech mode" and an "auditory mode". Despite attempts to dismiss this dichotomy, additional evidence continues to accumulate as has been suggested by several of the new findings summarized in the papers included in this symposium.

The picture is far from clear, however, because the problems inherent in comparing speech and nonspeech signals have generated several questions about the interpretation of results obtained in earlier studies. First, there is the question of whether the same psychophysical properties found in the speech stimuli were really preserved in the parallel set of nonspeech control signals. Such a criticism is appropriate for the original /do/--/to/ nonspeech control stimuli which were simply inverted patterns reproduced on the pattern playback. The same remarks also apply to the well-known "chirp" and "bleat" control stimuli of Mattingly et al. (1971) which were created by removing the formant transitions and steady-states from the original speech context. These stimuli were presented in isolation to subjects for discrimination. Such manipulations, while nominally preserving the phonetic "cue" obviously result in marked changes in the spectral context of the signal which no doubt affects the detection and discrimination of the original formant transition. Such criticisms have been taken into account in the more recent experiments comparing speech and nonspeech signals as summarized by Dr. Dorman and Dr. Liberman, in which the stimulus materials remain identical across different experimental manipulations. While these more recent studies relieve some of the ambiguities of the earlier experiments, problems still remain in drawing comparisons between speech and nonspeech signals. For example, subjects in these experiments rarely practice with the nonspeech control signals to develop the competence required to categorize them consistently. With complex multidimensional signals it is quite difficult for subjects to attend to the relevant attributes that distinguish one signal from others presented in the experiment. A subject's performance with these nonspeech signals may therefore be no better than chance if he/she is not attending selectively to the same specific criterial attributes that distinguished the original speech stimuli. Indeed, not knowing what to listen for may force a subject to attend selectively to an irrelevant or misleading attribute of the signal itself. Alternatively, a subject may simply focus on the most salient auditory quality of the perceived stimulus without regard for the less salient acoustic properties which often are the most important in speech such as burst spectra or formant transitions. Since almost all of the nonspeech experiments conducted in the past were

carried out without the use of discrimination training and feedback to subjects, an observer may simply focus on one aspect of the stimulus on one trial and an entirely different aspect of the stimulus on the next trial. Without training experience to help the subject identify the criterial properties, the observed performance may be close to chance, a result that has been reported quite consistently in the literature. Setting aside some of these criticisms, the question still remains whether drawing comparisons in perception between speech and nonspeech signals will yield meaningful insights into the perceptual mechanisms deployed in processing speech. In recent years, the use of cross-language, developmental and comparative (i.e., cross-species) designs in speech perception research has proven to be quite useful in this regard as a way of separating out the various roles that genetic predispositions and experience play in speech perception. On the other hand, these types of investigations provide needed information about the course of learning and perceptual development since spoken language must be acquired in the local environment through social contact. On the other hand, comparative studies with both speech and nonspeech stimuli are useful in defining the lower limits on auditory system function. However, there are serious limitations in studies of this kind. For example, while it is cited with increasing frequency that chinchillas categorize synthetic stimuli differing in VOT in a manner quite similar to English-speaking adults, little if anything is ever mentioned, however, about the chinchilla's failure to carry out the same task with stimuli differing in the cues to place of articulation in stops, a discrimination that even young prelinguistic infants can make (Eimas, 1974). Should we then conclude that the English voicing contrast is purely sensory in origin, while place of articulation or voicing in Thai is somehow more "linguistic", brought on by inheritance or very early experience? With a little reflection, I think the answer must surely be negative. Such comparative studies are useful in speech perception research but only to the extent that they can specify the lower-limits on the sensory properties of the stimuli themselves. However, these findings are incapable, in principle, of providing any further information about how these signals might be "interpreted" or coded within the context of the experience and history of the organism.

Animals simply do not have spoken language and they do not and cannot recognize, as far as I know, differences between phonetic and phonological structure, a fundamental dichotomy in all natural languages. Cross-language and developmental designs have also been quite useful in providing new information about the role of early experience in perceptual development and the manner in which selective modification or tuning of the perceptual system takes place. Although the linguistic experience and background of a listener was once thought to control his/her discriminative capacities in speech perception experiments, recent findings strongly suggest that the perceptual system has a good deal of plasticity for retuning and realignment, even into adulthood. The extent to which control over the productive abilities remains plastic is still a topic to be explored. To what extent is it then useful to argue for the existence of different modes of perception for speech and nonspeech signals? Some investigators such as Dr. Ades would simply dismiss the distinctions drawn from earlier work on the grounds of parsimony and generality. He has argued recently (Ades, 1977) and in his contribution to this symposium that differences in perception between speech and nonspeech or consonants and vowels can be accounted for simply by recourse to the notion of "range" or the width of the context expressed in terms of the number of JNDs. As long as the range is small, absolute identification performance will be as good as differential discrimination. When the range is large, however, discrimination will be better than identification. Thus according to the account offered by Ades, a consonant continuum should display a smaller range than a vowel continuum. But as shown in Fig. 1 the facts are quite the reverse of his predictions.

In this figure we have reproduced the identification data collected by Perey and Pisoni (1977) in a magnitude estimation task. On each trial subjects had to respond to a stimulus with a rating on a scale from 1 to 7. One group of subjects received a consonant continuum differing in VOT, another received a vowel continuum. Through various transformations of the obtained stimulus-response matrix, scale scores were derived and an estimate of the perceived psychological spacing between stimuli was obtained. Scale scores are expressed in this figure in terms of

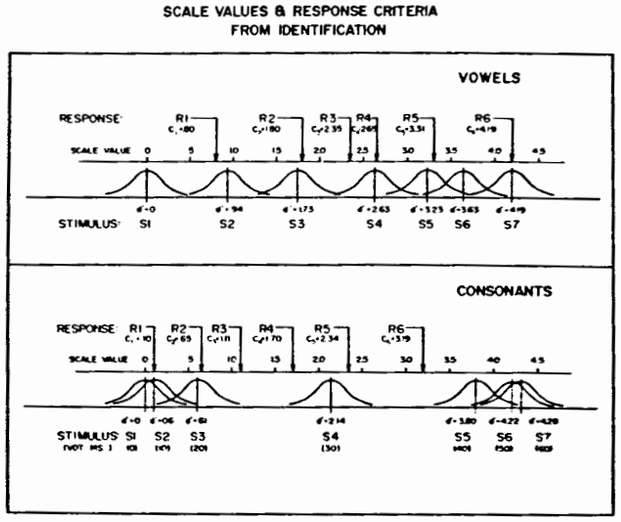


Figure 1. Scale values showing the perceived psychological space for consonants and vowels. Data were taken from Perey and Pisoni (1977) who required subjects to use a rating response in identification.

d's and by summing these individual values, an estimate of the total range or spacing of the stimuli was obtained. The cumulative d' is shown on the far right of each panel. Notice that the cumulative d' for the vowels shown on the top is 4.19 while the value for the consonants shown on the bottom is 4.28. If stimulus range were the correct explanation of the differences in perception between consonants and vowels as Dr. Ades would have us believe, the consonants should have displayed the smaller range. Obviously, this is simply not the case. However, what is of interest in this figure is the psychological spacing of signals within each panel. For the consonants, the spacing between adjacent stimuli is clearly unequal with a grouping close to the endpoints of the series. For the vowels, the spacing is more nearly equal across all the test stimuli suggesting the possibility of better resolution in discrimination, a result that has been known for

many years. Thus, Dr. Ades' argument that the range of stimuli can account for differences in perception between consonants and vowels or speech and nonspeech would seem to be incorrect, despite his attempts to generalize the Durlach and Braida (1969) model to speech perception. Moreover, this is a curious position to maintain anyway as it is commonly recognized, not only in speech perception research but in other areas of perceptual psychology, that "nominal" stimuli may receive differential amounts of processing or attention by the subject, that subjects may organize the interpretation of the sensory information differently under different conditions and that the sensory trace of the initial input signal may show only a faint resemblance to its final internal representation resulting from encoding and storage in memory. It is hard to deny that a speech signal elicits a characteristic mode of response in a human subject--a response that is not simply the consequence of an acoustic waveform leaving a meaningless sensory trace in the auditory periphery. Nevertheless, there is a great deal to learn about how the auditory system codes complex acoustic signals such as speech. Dr. Dorman, in summarizing work on the perception of transitions in speech and nonspeech context, has tried to establish the need for a specialized speech processor to account for differences in labelling of sine-wave stimuli when heard as either speech or nonspeech. Such explanations seem to me entirely premature at this time as the relevant psychophysical experiments with nonspeech signals have simply not been carried out yet. To remedy this state of affairs we have begun to collect labelling data in our laboratory recently using brief FM stimuli followed by a constant frequency (CF) steady-state. Schematized spectrograms of the test stimuli are shown in Fig. 2.

The left panel of this figure shows an idealized set of stimuli differing in the initial starting frequency of the FM. Three steady-state (CF) frequencies were selected, 850, 1500 and 2300 Hz. For each set we generated 21 test signals which spanned a range of 500 Hz above and below the CF of the steady-state component. In Experiment I the three sets of signals consisted of an isolated single component as shown on the left. In Experiment II we added an additional 500 Hz component to each of the original three sets of stimuli. Subjects were required to identify the

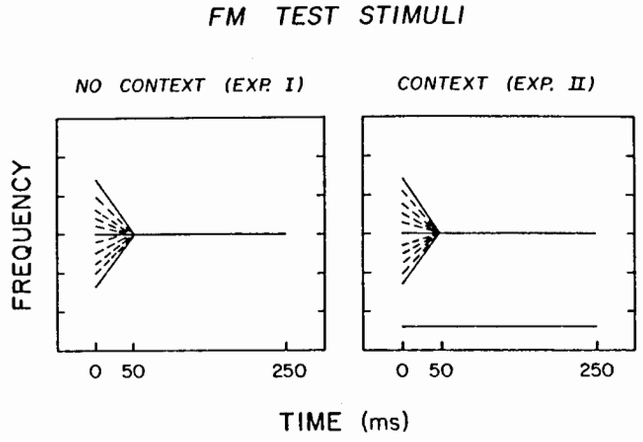


Figure 2. Schematized patterns showing the time course of the non-speech FM stimuli: The panel on the left illustrates the test stimuli without spectral context, the panel on the right shows the addition of a low frequency component to the same signals.

stimuli as "rising", "level" or "falling" after a brief training period with good exemplars selected from each category. The results of both experiments are shown in Fig. 3.

The labelling functions shown at the top for the three CF conditions reveal that the middle or "level" category response increases slightly in size as the CF of the steady-state increases from 850 Hz to 1500 Hz, a result that is consistent with what is known about frequency resolution in the auditory system. Over a wide range of frequencies, discrimination follows Weber's law. Thus, the level category should widen as the frequency of the steady-state increases for the same difference in initial starting frequency. Note that we have plotted starting frequency on a linear rather than log scale. The results for Experiment II in which an additional steady-state component was added are shown in the lower panel of the figure. Notice that for the 850 Hz condition the "level" category is now slightly larger than in the top panel suggesting the strong possibility of some interaction between the individual components. However, the other two condi-

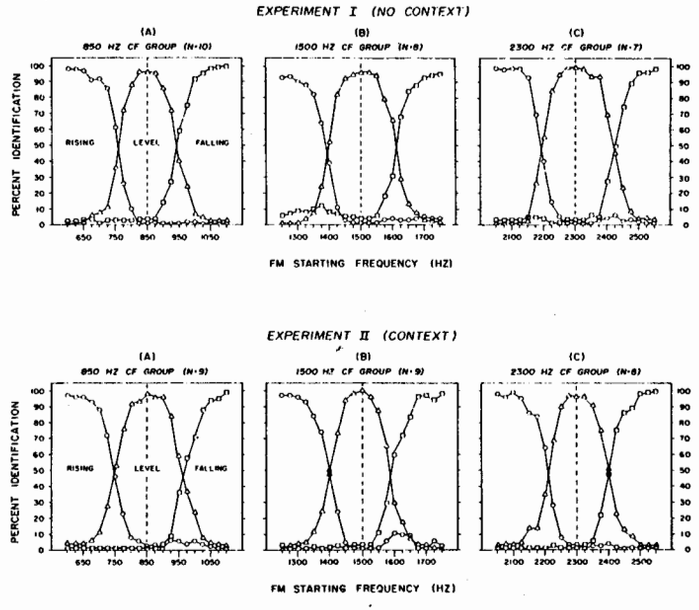


Figure 3. Identification data for FM stimuli obtained with three different steady-state CF's, 850 Hz, 1500 Hz and 2300 Hz. The top panel shows the identification data collected for FM's without context, the lower panel shows the data for test signals with the additional steady-state context present.

tions in Experiment II show a somewhat narrower range for the "level" category compared to the top panel indicating better resolution of frequency in the presence of another signal, a well known fact in auditory psychophysics. These recent findings were not originally intended to refute the arguments of Dorman and his colleagues who favor the postulation of some specialized perceptual mechanism for processing speech signals. Rather, I simply wanted to illustrate by way of example that the location of perceptual categories observed with nonspeech signals is not rigidly controlled by some simple physically defined invariant such as the direction of the frequency change. Moreover, as Dr. Divenyi has pointed out so well in his paper, we need to know much more about how the

basic constraints of the auditory system affect the way speech is initially coded for subsequent processing. Thus, in the present case several basic facts about frequency discrimination are sufficient to account for changes in our subjects' perceptual categorization of nonspeech FM's that are similar to speech. Whether it will be possible to generalize such psychophysical explanations to more complex signals such as speech remains to be seen from future research currently in progress in our laboratory and elsewhere.

In summary, there still appears to be good evidence for distinguishing between speech and nonspeech signals and for recognizing the existence of two distinct modes of perception, one associated with the sensory or psychophysical correlates of acoustic signals and the other with the interpretation and coding of acoustic signals as speech. Recent work has attempted to make these differences more precise by subjecting them to experimental test and searching for common underlying explanations. Taken together such results suggest to me that, just as in the case of "species-typical responding" observed in the behavior of other animals, the notion of a "speech mode" of perception captures certain aspects of the way human observers typically respond to speech signals that are highly familiar to them. We still do not know if it is simply a matter of familiarity as with music or whether there is something deeper and more closely related to biological survival of the organism. Nevertheless, such a conceptualization does not, at least in my view, commit one to the view that human listeners cannot respond to speech signals in other ways more closely correlated with the sensory or psychophysical attributes of the signals themselves. To deny the speech mode, however, is to ignore the fact that acoustic signals generated by the human vocal tract are used in a distinctive and quite systematic way by both talkers and listeners to communicate linguistically, a species-typical behavior that is restricted, as far as I know, to Homo sapiens.

Past experiments comparing the perception of speech and nonspeech signals have been quite useful in characterizing how the phonological systems of natural languages have, in some sense, made use of the general properties of sensory systems in selecting an inventory of phonetic features and their acoustic correlates (Stevens, 1972). The relatively small number of distinctive features and their acoustic correlates that can be observed across a wide variety of diverse languages implies that there is a common sensory basis for language perception, a common means of controlling the mechanisms of speech production and a common cognitive definition of linguistic structure. Whether these facts are causally related will no doubt be a matter of much debate, speculation and new research in the years to come. It is clear, nevertheless, that the distinctions drawn in perception between speech and nonspeech signals still remain fundamental, setting apart research on speech perception from the study of auditory psychophysics and the field of auditory perception more generally.

Acknowledgements

The preparation of this paper was supported, in part, by NIMH grant MH-24027 and NINCDS grant NS-12179 to Indiana University in Bloomington. I am grateful to Peter Jusczyk and Jim Sawusch for comments on an earlier draft of the paper. Robert Remez discussed many of the theoretical issues summarized in the paper with me at length and provided helpful editorial comments that improved the overall exposition and quality. His help is greatly appreciated.

References

Ades, A.E. (1977): "Vowels, consonants, speech and nonspeech", Psych. Rev. 84, 524-530.

Cutting, J.E. and B.S. Rosner (1974): "Categories and boundaries in speech and music", Perc. Psych. 16, 564-570.

Durlach, N.I. and L.D. Braida (1969): "Intensity perception I. Preliminary theory of intensity resolution", JASA 46, 372-383.

Eimas, P.D. (1974): "Auditory and linguistic processing of cues for place of articulation by infants", Perc. Psych. 16, 513-521.

Liberman, A.M., K.S. Harris, H.S. Hoffman, and B.C. Griffith (1957): "The discrimination of speech sounds within and across phoneme boundaries", J.Exp.Psych. 54, 358-368.

Liberman, A.M., K.S. Harris, J.A. Kinney, and H.L. Lane (1961): "The discrimination of relative onset time of the components of certain speech and non-speech patterns", J.Exp.Psych. 61, 379-388.

Mattingly, I.G., A.M. Liberman, A.K. Syrdal, and T.G. Halwes (1971): "Discrimination in speech and non-speech modes", Cogn.Psych. 2, 131-157.

Miller, J.D., C.C. Wier, R. Pastore, W.J. Kelly, and R.J. Dooling (1976): "Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception", JASA 60, 410-417.

Perey, A.J. and D.B. Pisoni (1977): "Dual processing versus response-limitation accounts of categorical perception: A reply to MacMillan, Kaplan and Creelman", JASA 62, S1, 60-61.

Pisoni, D.B. (1977): "Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stops", JASA 61, 1352-1361.

Stevens, K.N. (1972): "The quantal theory of speech: Evidence from articulatory-acoustic data", in Human communication: A unified view, E.E. David, Jr. and P.B. Denes (eds.), New York: McGraw-Hill.

COMMENTS FROM THE PANELISTS

The symposium on the perception of speech and nonspeech began with a brief summary statement by each of the contributors. This was followed by a panel discussion dealing with several issues that came up during the presentations. Finally, a number of questions and comments from the general audience were presented, followed by further discussion by the members of the panel. The highlights of these discussions and interactions are summarized below in an attempt to capture the flavor of the general issues and problems that surfaced as a result of this symposium.

Dr. Ades began his presentation by summarizing his paper contributed to the symposium and offering several comments on the introductory remarks given earlier by Professor Pisoni. Dr. Ades reiterated several times in this presentation that he personally believed that speech perception was, in some sense, "unique" or "special" despite the weak evidence usually cited from identification and discrimination experiments. He argued that the differences in perception between speech and nonspeech signals or consonants and vowels could be accounted for by differences in the range or spacing of signals. Dr. Ades criticized the recent data presented by Professor Pisoni showing equivalent ranges for consonants and vowels on the grounds that these data were collected in an identification rather than a discrimination paradigm. Most of Dr. Ades' specific remarks were directed, however, at narrow experimental questions, particularly the use of high uncertainty discrimination paradigms which provide relatively low estimates of discriminability.

Dr. Divenyi argued for the operation of two stages of processing in auditory perception regardless of whether the signals are complex auditory patterns or speech signals. According to Dr. Divenyi, speech is simply one class of complex signals with which the listener has had extensive experience and familiarity. Dr.

Divenyi described his two-stage model of auditory processing. The first stage, the auditory stage, involves the sensory analysis and coding of signals by the peripheral auditory system. The representation of signals at this stage is something like a neurogram reflecting the frequency selectivity of the auditory system. The second stage, the temporal stage in Dr. Divenyi's model, involves the analysis and coding of temporal information or patterns in both speech and nonspeech signals. Dr. Divenyi argued that the differences in perception between speech and nonspeech signals were due to differences in listening strategies brought about by learning and experience with speech and other sounds. Thus, in listening to speech several different strategies are available to the listener for centering or positioning the listening band differently. Dr. Divenyi concluded that there were no structural differences in perception between the so-called "speech mode" and "nonspeech modes" of processing. The distinctiveness of speech arises, according to Dr. Divenyi, from mere exposure and familiarity with speech and not because of any specialized processing by the auditory system.

Professor Dorman summarized his recent research which was carried out in collaboration with Drs. Bailey and Summerfield. This work was concerned with the perception of speech and nonspeech stimuli differing in the cues to place of articulation. Professor Dorman stated that his interest in these comparisons grew out of several questions surrounding whether infants can perceive speech signals as speech rather than simply complex nonspeech patterns. The methodology employed in these studies using adult subjects involved comparisons dissociating the location of the "phonetic" boundary from the location of the "acoustic" boundary. The results of these tests showed differences in the loci of the boundaries depending on whether the nonspeech stimuli were heard as speech or nonspeech. Accordingly, Professor Dorman argued for the operation of two modes of processing nonspeech signals having speech-like properties. Furthermore, Professor Dorman implied that the dissociation of these two modes could be assessed by looking at differences in the location of category boundaries when the same stimuli are perceived as speech or nonspeech.

Professor Massaro departed from his symposium contribution by focusing on his general model of auditory information processing which postulates both structures for storage of information in memory and processes for carrying out various operations on this information. According to Professor Massaro's model, the earliest stage of processing involves acoustic feature analysis and is similar for speech and nonspeech signals alike. Processing here is not influenced by higher-order knowledge or context from long-term memory. Professor Massaro claimed that his general model could account for the differences observed in perception between speech and nonspeech without assuming the existence or operation of a specialized "speech mode" of processing. According to Professor Massaro, a listener's higher-order knowledge and his experience with speech affects the way acoustic features are treated and integrated at what he calls the primary stage of recognition in his model. Thus, a two stage model is also assumed to be necessary for perception of speech stimuli although the same two processes may be employed with other nonspeech stimuli.

Professor Liberman's remarks on duplex perception were summarized very briefly by Professor Studdert-Kennedy.[1] Using a variation of the so-called "Rand Effect", Professor Liberman has shown that listeners can simultaneously perceive a phonetic event (i.e., a CV syllable) and an auditory event (i.e., a chirp). Professor Liberman has argued that these results imply that both auditory and phonetic processes are carried out together simultaneously in parallel and that a distinct phonetic subsystem exists for processing speech signals, a subsystem which is separate from processes used to perceive other auditory signals.

Dr. Summerfield summarized his symposium paper with Dr. Bailey by emphasizing that the information for phonetic perception must be found in the acoustic signal itself which reflects the consequences of articulation of speech. Dr. Summerfield suggested that the phonetic information in the signal could be properly characterized by detailed examination of the articulatory control that gives rise to acoustic patterning in speech production and by a detailed examination of how the distinctiveness of this articulatory patterning is enhanced by auditory processing of speech signals.

---

1) Professor Liberman was not present at the congress.

Dr. Summerfield emphasized that this research strategy would be possible without having to assume any need for articulatory mediation in speech perception.

DISCUSSION

Following the individual summary statements, there was a general discussion among the panel members which was then opened up to the audience for additional questions and comments. Several broad and narrow issues appeared to emerge from the symposium papers and summary presentations as well as from the preliminary discussions that the panel members held before the symposium began.

Professor Studdert-Kennedy summarized these issues briefly before beginning the panel discussion. The first, and perhaps most general issue, concerned comparisons made in perception between speech and nonspeech signals. Specifically, it appeared that everyone agreed more or less that speech perception is in some sense special although not everyone agreed on precisely in what way it is special. Thus, the question of whether speech is a special process is one that still remains and apparently is one that continues to occupy the attention of numerous investigators working in speech perception even today.

Closely associated with the speech-is-special issue is a set of somewhat more narrowly defined experimental issues related to how one would be able to demonstrate clearly what the presumed special properties of speech are. That is, some concern was expressed among several members of the panel with the currently available methods and research paradigms used in speech perception research, particularly the use of discrimination procedures to assess differences between speech and nonspeech signals. During Dr. Ades' summary statement and later during the panel discussion, he repeated his dissatisfaction and skepticism with the traditional methods of comparing identification and discrimination of speech and nonspeech and consonants and vowels.

Another, somewhat broader issue that emerged from these discussions concerned the question raised by Summerfield and Bailey in their paper of whether there are, in fact, "characteristic" acoustic properties of speech signals that result directly from articulation and whether these properties are distinct from the properties of nonspeech signals. This particular issue highlights

the clear separation of views that emerged at the symposium by Divenyi and Massaro, for example, who suppose instead that there really are no distinctively different or unique acoustic correlates of speech sounds that separate them from the class of nonspeech signals in the listener's environment. According to both of these investigators, differential processing by a human observer is not required or determined by properties of the signal itself but rather by experience, training, context and higher-order knowledge. The early stages of perceptual processing are therefore the same for speech and nonspeech signals alike.

Finally, the issues surrounding the development of speech perception, particularly the recent findings with young prelinguistic infants, were also cited as a potentially important topic for further discussion. Professor Studdert-Kennedy wondered to what extent it is reasonable to suppose that an organism such as a young infant who does not "know" a language can respond to an acoustic signal as though it were conveying language--that is as though the signal were speech.

The panel discussion began with several additional remarks about the use of discrimination paradigms in speech perception research. Dr. Ades suggested that he could see little use for additional discrimination experiments in the future. Dr. Divenyi repeated several of his earlier comments on the need for two stages of processing in auditory perception to deal with all the relevant empirical phenomena in the literature. Moreover, he restated his claims again about the role of perceptual strategies in determining what a listener focuses his attention on in speech perception.

In responding to Dr. Ades' remarks about discrimination testing, Dr. Massaro felt that discrimination experiments should proceed in parallel with categorization experiments to illuminate the nature of processing speech and nonspeech. Moreover, Dr. Massaro summarized the results of recent experiments that manipulated several acoustic cues at the same time in order to explore how listeners integrate or combine information in complex multidimensional signals.

Professor Studdert-Kennedy suggested that the discussion seemed to point toward general agreement about the need for levels and stages of processing in perception, particularly speech perception. Professor Studdert-Kennedy also noted at this time that

one of the major reasons for postulating two levels in speech perception was the earlier work of Fujisaki suggesting the possibility that two kinds of auditory memory or coding were operating in categorical perception experiments.

The discussion then turned to the issue of how speech is distinguished acoustically from nonspeech signals. Dr. Summerfield pointed out that the contrast between speech and nonspeech might be more profitably examined in terms of different styles of processing--one appropriate for real world "events" (i.e., speech signals generated by a human vocal tract) and the other being appropriate for a relatively unnatural mode of processing where the object of interest is a "nonevent". Dr. Summerfield also suggested that there are reliable acoustic markers in the speech signal that inform a listener that the signal is speech rather than nonspeech. For example, the posture of the vocal apparatus during speech production is unique to speaking. There are both short- and long-term changes in variations in intensity and rise-time which are indicators of speech that may act as "trigger-features" to engage a speech mode of processing.

Professor Dorman then suggested a possible experimental paradigm to compare speech and nonspeech more directly by examination of "trading relations" between different types of acoustic cues in both contexts. If the trading relations differ between the two contexts, speech and nonspeech, then one could argue for distinctly different modes of processing for speech vs. nonspeech signals.

After the members of the symposium panel completed their discussion of these issues, the moderator opened the discussion to members of the general audience in attendance. Professor Stevens raised the issue again of what markers or characteristics distinguish speech from nonspeech signals. Professor Stevens suggested that it is not necessary to make reference to articulation in speech perception because all speech signals have three or four criterial acoustic properties that set them apart from all nonspeech signals. The first property involves the rate of amplitude variations over time. A basic property of speech is that it has a syllabic structure creating amplitude fluctuations between consonants and vowels. A second property of speech is shown in the spectra of speech signals. If the spectra of speech are sampled

at any point in time, the resulting analysis will display characteristic peaks and valleys. A third property of speech is the fact that these spectra change with time. That is, there are well-defined acoustic correlates to the changing articulatory gestures in speech production. The spectra of speech can also change rapidly or slowly over time. Professor Stevens suggested that one might speculate that speech signals are acoustic signals that the auditory system "likes" because it is easy to extract properties from signals of this kind.

Dr. Waterson then raised the question of the usefulness of the present kinds of experiments carried out on speech vs. nonspeech. She argued that almost all of the research has used European-based languages with either European or American subjects and the tests employ language-specific features such as VOT. That is, the contrasts are presented in the language of the subjects. She wondered what sorts of results would be obtained if the subjects were presented with sounds from more exotic languages.

Professor Kuhl questioned the claim made earlier in the introduction by Professor Pisoni concerning the chinchilla's apparent inability to discriminate some of the cues to place of articulation in stop consonants. Professor Kuhl pointed out that the chinchilla's failure to discriminate /d/ from /g/ is due to a basic sensory limitation involving the length of their basilar membrane and not any inherent perceptual or cognitive limitation. Professor Kuhl also took issue with another remark of Professor Pisoni's in his introduction concerning the usefulness of certain kinds of comparative designs involving animal subjects and what these results could provide for understanding human language. Professor Kuhl stated that very pertinent information about "processing" species-specific acoustic signals may be provided by looking at animal models, particularly animals in which "vocal learning" is a salient characteristic such as the acquisition of bird song or coos by certain species of monkeys. Unfortunately, Professor Kuhl did not provide any further details about precisely what kinds of information would be obtained from these animal studies nor how the perceptual processing by these animals could be compared to the analyses carried out by humans.

Professor Kuhl also touched on the issue of a predisposition for processing certain salient acoustic attributes by human infants.

Such salient properties might serve to "focus" the infants' attention on certain aspects of the speech signal at a very early age. Moreover, Professor Kuhl repeated the suggestion, made by several others, that there is the strong possibility that the selection of speech sounds in language was guided, in some sense, by evolutionary constraints on the close match between both speech production and speech perception.

Dr. Klatt pointed out an important methodological difference in the results presented in the introduction by Professor Pisoni and the findings obtained by Professor Dorman on sine-wave analogs of CV syllables. Professor Pisoni showed well-defined labelling data for three categories of FMs corresponding to rising, level and falling, whereas Professor Dorman only reported two categories corresponding to rising and falling. Dr. Klatt suggested that this is a potentially important issue worthy of further study with fine-grained discrimination techniques which reduce the use of category labels. Dr. Klatt raised the question again of whether speech signals are somehow structured along "natural" auditory or psychophysical distinctions and/or constraints from the way speech is produced by the articulatory system.

Professor Fourcin offered an additional property, variations in fundamental frequency, that should be added to Professor Stevens' list for distinguishing between speech and nonspeech signals. Professor Fourcin also emphasized the need to look at pattern learning as the abstraction of invariants in complex stimuli, a topic that received little, if any, attention by members of the symposium.

Following the questions and comments from the audience, each of the panel members provided several additional final remarks elaborating on the statements they made earlier or commenting on some specific item raised in the general discussion. For the most part, however, the symposium on speech vs. nonspeech served to solidify a general sense of agreement among various investigators as to the value of comparisons in perception between speech and nonspeech signals. The issue of whether speech is special was discussed extensively throughout the symposium and led to a consensus that such a broad distinction is no longer meaningful, although nearly everyone believed that speech perception was somehow special or unique in its own way. A central issue that emerged

from this symposium was a concern with identifying the distinctive acoustic properties of speech signals that set them apart from other nonspeech signals in the listener's environment. There was also some attention devoted to questions of perceptual development in infants and issues surrounding perceptual predispositions for processing speech signals. Finally, there was a continued lively debate and interaction throughout the symposium on research methodology, particularly the use of discrimination paradigms in speech perception and the relevance of these sorts of data to categorization and recognition of phonemes in speech.

WORKING GROUP: THE SYLLABLE IN PHONOLOGICAL THEORY

Organizer: Alan Bell

ALAN BELL's SUMMARY

The Working Group met twice during the Congress to discuss selected issues related to the controversial unit of phonetics and phonology. The discussions largely concerned questions raised by the following papers, which the authors had exchanged among themselves and a few other researchers before the congress.[1]

Árnason, Kristján: "A diachronic look at the syllable"

Bell, Alan: "The syllable as a constituent versus organizational unit"

Bell, Alan: "The role of segment bonds in phonological organization"

Brend, Ruth: "The syllable in tagmemic analysis"

Coates, Richard: "A point of universal phonotactics?"

Coates, Richard: "The categories of real phonology in relation to the syllable"

Coates, Richard: "Some allegro syllabic consonant processes in English"

Coates, Richard: "Reservations on the origin of syllabic consonants"

Cochran, Anne M.: "Notes on current research on the syllable in Papua New Guinea languages"

Cochran, Anne M.: "Ampeeli-Wojokeso consonant clusters--a study in syllable complexity" (with Edith and Dorothy West)

Galton, Herbert: "Interrelations between the open syllable and the phonological system as illustrated in Slavic"

Mikuš, Radivoj: "Vers une nouvelle phonétique"[2]

Price, Patti Jo: "What is the syllable anyway?"

The workshop was also fortunate to have the participation of the following Congress attendees with research experience on the syllable and related matters: H. Andersen, B. Andrésen, C.-J.N. Bailey, R. Bannert, H. Basbøll, R.A.W. Bladon, J. Bybee Hooper, W. Dressler, O. Fujimura, J. Gvozdanović, J.T. Jensen, C.-W. Kim, I. Lehiste, B. Lindblom, L. Menn, L. Papademetre, E. Pike, L. Selkirk, E. Strangert, S. Vater and K. Williamson.

---

1) Requests for copies of papers should be addressed to the individual authors.

2) R. Mikuš was unfortunately not able to attend the Congress.

The first session opened with discussion of Price's experiments on the acoustic cues sufficient to shift identification of tokens prepared with the aid of speech synthesis among prayed-parade-braid-bereted and among plight-polite-blight-belight. This led to a general discussion of a wide variety of such phenomena, including some of particular interest mentioned in Cochran's paper, and of acoustic cues involved. Some comment on the different ways judgements on the number of syllables can be obtained also followed. Discussion then turned to the concept of the relative "resistance to coarticulation" of segment classes presented by Lindblom and Bladon and to a theory of the internal structure of the syllable sketched by Basbøll. The session concluded with discussion of Coates' proposal that the syllable functions as a domain of feature timing in a phonological theory in which time rather than sequence is the basis of phonological representation.

The first topic of the second session was the role of the syllable in diachronic phonology, under which three cases were taken up. These were Galton's contention that the open syllable canon of Slavic was a principal factor in the development of the correlation of palatalization, Árnason's study of vowel shortening and lengthening in Icelandic, which he concluded to be inadequately explained by several different theories of syllabic representation, and the case of cluster formation in Modern Greek presented by Papademetre. The final topics of the workshop were Bell's proposed framework of segment bonding as an alternative to current syllabic models and the general question of the hierarchical nature of the syllable as described in tagmemic theory by Brend.