# WORKING PAPERS IN LINGUISTICS NO. 6

by

Charles J. Fillmore, Ilse Lehiste,
David Meltzer, Marcel A. A. Tatham, and
Sandra Annear Thompson

TECHNICAL REPORT NO. 70-12

September, 1970

# Foreword

The Computer and Information Science Research Center of The Ohio State University is an inter-disciplinary research organization which consists of the staff, graduate students, and faculty of many University departments and laboratories. This report presents research accomplished in cooperation with the Department of Linguistics.

The work of the Center is largely supported by government contracts and grants. The preparation of the papers contained in this report was partly supported by the Office of Science Information Service, National Science Foundation under Grant No. GN-534.

Ilse Lehiste

Ilse Lehiste
Chairman
Department of Linguistics

Marshall C. Yovits
Director
CIS Research Center

## Table of Contents

## List of WORKING PAPERS IN LINGUISTICS

No. 1

"The Grammar of 'Hitting' and 'Breaking'," Charles J. Fillmore, pp. 9-29. (To appear in Studies in English Transformational Grammar, R. Jacobs and P. Rosenbaum, eds., Ginn-Blaisdell, 1970.)

"The English Preposition WITH," P. Gregory Lee, pp. 30-79.

"Relative Clauses and Conjunctions," Sandra Annear Thompson, pp. 80-99.

"On Selection, Projection, Meaning, and Semantic Content," D. Terence Langendoen, pp. 100-109.

"Some Problems of Derivational Morphology," Sandra Annear Thompson and Dale Elliott, pp. 110-115.

"The Accessibility of Deep (Semantic) Structures," D. Terence Langendoen, pp. 118-127. (To appear in Studies in English Transformational Grammar, R. Jacobs and P. Rosenbaum, eds., Ginn-Blaisdell, 1970).

"Review of Haim Gaifman, 'Dependency Systems and Phrase-Structure Systems,' Information and Control 8 (1965), pp. 304-337," James T. Heringer, pp. 128-136.

"Diphthongs Versus Vowel Sequences in Estonian," Ilse Lehiste, pp. 138-148. (To appear in the Proceedings of the VI International Congress of Phonetic Sciences, Prague, 1967.)

No. 2

"Lexical Entries for Verbs," Charles J. Fillmore, pp. 1-29. (Also in *Foundations of Language* 4 (1968), pp. 373-393.)

"Review of *Componential Analysis of General Vocabulary: The Semantic Structure of a Set of Verbs in English, Hindi, and Japanese*, Part II, by Edward Herman Bendix. I.J.A.L. Vol. 32, No. 2, Publication 41, 1966." Charles J. Fillmore, pp. 30-64. (Also in *General Linguistics* 9.41-65 (1969)).

"Types of Lexical Information," Charles J. Fillmore, pp. 65-103. (To appear in *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics, Anthropology and Psychology*, Jacobovits and Steinberg, eds., Cambridge University Press; and *Proceedings of the Balatonszabadi Conference on Mathematical Linguistics*, Kiefer, ed., D. Reidel.)

"'Being' and 'Having' in Estonian," Ilse Lehiste, pp. 104-128. (Also in *Foundations of Language* 5 (1969), pp. 324-341.)

No. 3

"*Do* from *Occur*," P. Gregory Lee, pp. 1-21.

"The Syntax of the Verb 'Happen'," Dale E. Elliott, pp. 22-35.

"Subjects and Agents," P. Gregory Lee, pp. 36-113.

"Modal Auxiliaries in Infinitive Clauses in English," D. Terence Langendoen, pp. 114-121.

"Some Problems in the Description of English Accentuation," D. Terence Langendoen, pp. 122-142.

No. 3 (continued)

No. 4

No. 4 (continued)

"A Double Regularity in the Acquisition of English Verb Morphology," Arnold M. Zwicky, pp. 142-148.

"An Annotated Bibliography on the Acquisition of English Verbal Morphology," Mary Louise Edwards, pp. 149-164.

No. 5

Twana Phonology, Gaberell Drachman, pp. i-286, Ph. D. dissertation, University of Chicago, 1969. [Limited printing; not sent out to everyone on the mailing list.]

# Introduction

This report deals with two aspects of research conducted in the Department of Linguistics under Grant No. GN-534, from the National Science Foundation. One of the aspects concerns itself with (i) the form and information content of lexical entries in a transformational grammar of English, (ii) the mechanism for lexical insertion in such a grammar, and (iii) the nature of the syntactic apparatus in terms of whose operations the lexical insertion process and the treatment of lexical information is to be carried through. The doctoral dissertation of S. A. Thompson was completed under this project; the present volume contains two articles based on the dissertation. Fillmore's paper on grammaticality deals with related issues.

The second aspect deals with communication by spoken language. During the past year, major efforts have been made to build up a facility for speech synthesis. This volume contains a detailed progress report (paper by David Meltzer). In connection with the building up of facilities, we have begun to identify research problems for which the synthesizer will serve as a tool. The goal is the testing of hypotheses about speech production and perception. A survey of various production and perception models has been made. Concrete results include the paper by M. A. A. Tatham on speech production models, and the translation (from the Russian, by I.

Lehiste) of a monograph by Bondarko et al., presenting a new model for speech perception.

# On Generativity*

Charles J. Fillmore

*Sponsored in part by the National Science Foundation through Grant GN-534.1 from the Office of Science Information Service to the Computer and Information Science Research Center, The Ohio State University.

# On Generativity*

Charles J. Fillmore

1. For some time I have been striving to understand just exactly what it takes for something to be a generative grammar. The nature of my concern with this question is not that of a metatheoretician within the discipline, nor that of a philosopher of science looking at our field from the outside; it is rather that of an easily confused Ordinary Working Grammarian who is trying to be minimally clear about what it is that he is doing.

The ordinary working grammarian of whom I speak has fairly special and fairly limited ways of troubling himself with the problems I will be discussing, and he has special and limited reasons for being pleased or displeased with a theory. For example, when the ordinary working grammarian is told that a generative grammar of a language is a recursive device which demarcates exhaustively and exclusively the unlimitedly large set of sentences in the language, what that means to him is that the theory gives him a test for knowing whether what he has done, in describing a certain language, has been successful: if he discovers sentences in the language which his grammar fails to recognize, or if he notices sequences which his grammar allows but the language does not, then he knows that his efforts have fallen short of complete success.

If the ordinary working grammarian is told that he can capture generalizations that would otherwise escape him only by adopting a particular notation or a particular set of conventions regarding the form and interpretation of grammatical rules, what that means to him is that the grammatical descriptions he writes should be simpler if he uses these notations and conventions than if he does not, and that grammars written by people who adhere to the same conventions will be interpretable to him.

Further, when the ordinary working grammarian is told that the model of grammar with which he should work must contain in its notation or in an auxiliary set of conventions a body of assumptions about language universals, he is willing to accept this, not so much because he is pleased that in this way the theory abstracts properties of the basic human psychic apparatus

*An earlier version of this paper was in need of surgery in several places, and the surgery was performed. I am grateful to the diagnosticians Arnold Zwicky and Stanley Peters.

for language out of the cultural diversity of individual languages, but because this decision makes it possible for him not to have to remember all the things he believes to be true about language in general: to the extent that his beliefs about language universals are embedded in the notations he uses, he will always know when to be surprised by new evidence which contradicts one or another of these beliefs. He knows that when he encounters linguistic facts which he cannot articulate with the notational and conceptual apparatus at his disposal, he has correctly detected a crisis in the theory and is now in a position to revise his beliefs about language.

Our grammarian, we have seen, is essentially lazy, and, indeed, almost 'practical' in his views about what theories are for.

I am going to claim that the ordinary working grammarian is confused about what it takes for something to be a generative grammar. Before I go on to explain myself, I must report immediately that we do not find him guilty of the much-discussed confusion between 'generate' as a stative verb used to relate a grammar and the sentences of the language it is a grammar of, and 'generate' as an active verb used of a human being and the utterances he produces. The ordinary working grammarian knows and is careful about these distinctions.[1]

I must also explain, before I go on, that the ordinary working grammarian I have in mind finds himself fairly solidly within the generativist camp. His doubts about generative grammar do not arise from any assumptions about the superiority of the research goals of the taxonomists or distributionists of a decade or two ago. To him, the data do not determine the conceptual base of the theory; they constitute, rather, the phenomena which the theory has to explain. And this was something he learned from the generativists.

For the sake of the younger reader, let me interpret my allusions. I am old enough to remember the days when, as a typical

[1] It is not so easy to keep these notions distinct in one's unconscious, I must admit. I continually find that I am attracted to what is called 'generative semantics' or back again to 'interpretive semantics' depending on whether I have recently been more impressed with my experiences of wanting to say things I do not know how to express, or with my experiences of having said things which I cannot understand. In the former mood I am convinced that the mechanism inside me for constructing well-formed messages is intact, and that what is malfunctioning is the component which maps messages into utterances; when I am in the latter state I feel that the mechanism for producing grammatical sentences is intact, and that what is defective is the apparatus for assigning meanings to them.

classroom demonstration of analytic procedures in linguistics, the professor presented a pair of linguistic forms, demonstrated on the basis of the distribution of their constituent elements that they are analogously constructed, and then continued by pointing out that their external distribution shows them to be distinct. I contribute the following examples for illustration: the pair 'maternity dress' and 'paternity suit'. It is easy to believe that there are distributional parallels in English-language texts between 'maternity' and 'paternity', and that the distributional properties of 'dress' and 'suit' are analogous. However, on examining the external distribution of the two-word expressions, we would discover that they are in fact quite distinct, in that they occur in vastly unlike total context sets. Some of my teachers took the trouble to say that when a linguist claims that two forms are grammatically distinct, all he means, in fact, is that their total context sets are distinct.

Today reasonable people are much more likely to say that there is something about what these expressions _are_ which accounts for their different distributions, rather than the other way around; and such reasonable people might be said to be taking the generativist position. To the challenge that these two ways of talking about the facts amount to the same thing, I reply that in the development of a generative description, one would notice the internal similarity of 'maternity dress' and 'paternity suit' only by accident; in the development of a distributionist account, the comparison of these forms is a necessary step in their individual description.

2. My topic, then, is the way in which a 'generative' linguist conceives the relation between a grammar and the objects which the grammar is designed to identify and describe, i.e., the 'grammatical' sentences of the language in question.

In the earliest discussions of generative grammars, a comparison was suggested between writing a grammar and specifying the set of well-formed formulas in a mathematical system. In Chomsky (1957, p. 13) we read, "The fundamental aim in the linguistic analysis of a language L is to separate the 'grammatical' sequences which are sentences of L from the 'ungrammatical' sequences which are not sentences of L and to study the structure of the grammatical sequences. The grammar of L will thus be a device that generates all of the grammatical sequences of L and none of the ungrammatical ones." A generative grammar recognizes certain strings of symbols as well-formed sentences in the language, but not others, much in the manner of the formation rules in a mathematical system.

This function of a grammar is interpretable as being identical to one of the unarticulated goals of the traditional grammarians, the difference being that a generative grammar is one in which the characterization of the totality of well-formed sentences is made explicit. To mention an aspect of such a suggestion which comes quickly to mind, it seems quite likely that some traditional grammarians, and many classroom grammarians, may indeed have been willing to think of a grammar as analogous to the system of formation rules in a mathematical system--that is, in the quite literal sense that in both cases the rules were devised by wise and rational creators, for the creators' own purposes, and that the admission or rejection of a presented formula or sentence was to depend on whether or not it was in conformity with these independently valued rules. A mathematical system and a system of grammatical rules upheld by proponents of the doctrine of correctness are both, after all, man-made.

Explicit generative grammars appeared on the scene, fortunately, at a time when the question of the membership of a sentence in a language was taken as an empirical issue. On the _de facto_, as opposed to the _de jure_, theory of grammaticality, the speaker is the source of the language, and a successful generative grammar is one which conforms in its predictions to certain kinds of judgments made by speakers of a language about the sentences in their language. A proposed grammar can be shown to be incorrect by a demonstration that the set of sentences in the language is not the same as the set of sentences recognized by the grammar.

That, at least, was the goal which grammarians learned to set for themselves. In the face of this first requirement, it is clear that what the ordinary working grammarian needs to _find_ out is the identity of the set of _de facto_ grammatical sentences, and what he needs to _figure_ out is whether the grammar he constructs puts the good sentences in and rules the bad ones out. We will see soon that this requirement is a difficult one.

In addition to the requirement that a grammar identify each of the grammatical sentences of the language, the concept of generative grammar comprises the further condition that it associate with each of the sentences it generates a structural description--a display of all of the grammatical information about the sentence which the speakers of the language can be said to possess. Our first two requirements are phrased in Katz (1966, p. 123) as follows: "the rules of a linguistic description must not only be capable of producing an infinite list of formal objects, but the formal objects on the list must be the sentences of the language under study and the list must exclude any string in the vocabulary of the language that is not a sentence in the language. Furthermore, these rules must somehow specify all the information about the sentences that a speaker utilizes to produce and understand them."

The second requirement does not commit us to anything new in the actual workings of a grammar. The very rules which play a part in the successful generation of the sentences of the language can be used, _via_ a structure-assigning algorithm taken to be part of linguistic theory, to provide the correct structural descriptions. As stated in Thorne (1968, p. 302), "The set of rules involved in the generation of a sentence is equivalent to an analysis of it."

With the concept of generative grammars thus elaborated to contain the notion 'correct structural description', the relation between grammar and the set of linguistic objects it generates is subtler than was apparent at first. The native-speaker judgments to which the analyst needs to appeal for convincing himself that his work is adequate involve not only acceptance or rejection of sentences, but also assent to various kinds of assertions about the sentences that are accepted.

Our ordinary working grammarian looks at this new responsibility and sees two problems: first, whether he can determine what the correct structural descriptions of the sentences in the languages are, and second, whether the rules needed for generating the sentences in the first sense are indeed precisely those which will succeed in assigning correct structural descriptions. The ordinary working grammarian worries, in other words, about whether there is really a definitional relation between a description of everything speakers know about the sentences of their language and grammatical rules of the type he has learned.

From the beginning, but only with seriousness in work later than Chomsky (1957), the concept of generative grammar has been further enriched by the requirement that it be capable of ranking sentences along a dimension ranging from the fully grammatical to the totally unstructured. It was apparently believed by Chomsky that for this new role there need be no new requirements on the form and operation of the generative apparatus itself. In Chomsky and Miller (1963, p. 291) we read that a generative grammar, *defined as* a device which enumerates the grammatical sentences of a language and which assigns structural descriptions to each of these, may also be *regarded as* a device which assigns to any string presented to it a relative-grammaticality index. What is needed, apparently, is some system of conventions which governs the way in which the structure-assigning apparatus is to be consulted for determining, for any non-sentence, its degree of departure from full grammaticality.

The ordinary working grammarian, confronting this added responsibility, sees now three things to worry about. The first is whether he or anyone he trusts knows how to rank sentences according to their degree of deviation from full grammaticality; the second is whether there is a general way of determining, from the rules of the grammar, a ranking of sentences which conforms to these judgments. His third problem is that he fails to understand why knowing what is wrong with each of two sentences should entail knowing whether one of them is worse off than the other.

One final enrichment of the concept of generative grammar is found in the view that a grammar which a grammarian constructs is a claim about something which speakers of the language have inside their skins and which makes them able to produce and comprehend the sentences, and many of the near-sentences, of their language (see Chomsky (1965, pp. 3-9)). With this addition the study of grammar takes on a new interest and importance, naturally, but with this addition one finds it particularly difficult to imagine in advance the precise nature of criteria for success. It will be my conclusion, nevertheless, that the most intelligible view of grammatical research sees it as the attempt to discover the internal rules which account for the rule-guided aspect of human linguistic abilities.

3. The most simply conceived goal of a generative grammar, to go back to the beginning, is that of determining, for any sequence of elements in the vocabulary of the language, whether it is grammatical or ungrammatical.

The details of the technical side of this task are of little real interest to the ordinary working grammarian. He knows that to the extent that any genuine generative grammar is an effective theory, it will always be possible to tell, _if_ a sentence is generated by the grammar, _that_ it is generated by the grammar: one tries out the rules, using whatever heuristic one has at hand, until one finds the sentence in question, and declares that it is in. There is, to be sure, another issue--that of knowing for certain that a presented string is _un_grammatical according to the grammar--but that question is related to subtle properties of grammars that are of little concern to the ordinary working grammarian. He is willing to assume that an interpreter of a generative grammar, given wit, luck and patience, will be able to find out one way or another whether a given sentence is in or out.

What does concern him is the non-technical problem of knowing whether the sentences that get in are the good ones and whether the sentences that get left out are the bad ones--whether, in other words, the grammar and the speakers make the same choices. He sees this as a problem because he knows that judgments about grammaticality are subject to all sorts of confusions between grammaticality and significance, acceptability or intelligibility; he knows that even when speakers say they understand that they are to make judgments about grammaticality rather than these other things, they still disagree; he knows that sometimes people change their minds about whether a sentence is grammatical; and he finds appeals to unending idiolectal variation somewhat unsatisfying.

There was a time when these uncertainties would not have bothered our grammarian: a decade ago there was little reason to doubt the Clear Cases Principle proclaimed in Chomsky (1957, pp. 13-14). On this principle, native-speaker judgments are criteria of grammar-constructing success only with respect to the clear cases. The grammarian begins by considering sentences like "I like ice-cream" that are clearly grammatical and sequences like "Ice-cream me the" that are clearly ungrammatical, and he constructs the simplest grammar which generates all the incontrovertibly grammatical sentences and fails to generate all the incontrovertibly ungrammatical sentences. The grammar, then, and not the grammarian, makes the decision about the unclear cases.

Today's grammarian finds little comfort in this principle, because he knows, if he has read Ross's thesis (Ross 1967), that the kinds of arguments that seem to bear very crucially on the nature and operation of syntactic systems involve him in grammaticality decisions that are extremely difficult to make. If he has seen the Elliot, Legum and Thompson (1969) studies of speech variation, he knows that properties of grammars and sentence configurations figure importantly in the description of idiolectal and stylistic differences, but not at all in a way that gives any primacy to a simple distinction between being in the language or out, being generated or not generated by the grammar.

The simplest criterion of success, which was to consist of checking the identity between being 'in the language' and being 'generated by the grammar', does not do, in short, what our ordinary working grammarian had hoped it would do for him.

4. But let us turn to another problem, that of designing a grammar capable of assigning degrees of grammaticalness. Chomsky's theory of relative grammaticality (see Chomsky 1965, pp. 148-154) takes roughly the following form. The grammar generates the set of fully grammatical sentences in a more or less straightforward way. For a string of words not found among the fully grammatical sentences, its degree of deviation from full grammaticality can be computed by comparing it with the grammatical sentences to which it is in some ways similar.

The procedure may be thought of as including something like the following steps. For each deviant string one identifies the set of sentences maximally similar to it. One identifies the properties which the deviant and the grammatical sentences have in common and in doing that one isolates just those properties which are 'out of place'. If an 'out-of-place' element is a constituent of a major category not found in that position in the grammatical sentences, the deviation is particularly serious--we may say that the string loses three points. Where an out-of-place element is of an appropriate category but has grammatical properties not found in that position in any of the fully grammatical sentences, the deviation is of minimal seriousness--the string loses one point. Where an out-of-place element is of an appropriate major category according to part of its context but requires ordinarily a categorial environment of a type not found in the string in question, the offense is of medium seriousness--the string loses two points. The degree of deviance of the string as a whole might be registered, in the most simple-minded rendering of this procedure, as the sum of the values of these various offenses.

The deviance-computing procedure I have just sketched, as well as subtler variations on it, has to be based on the assumption that it is in principle possible to identify, for a deviant string, just those lexical items or features which are out of place, or just those orderings of elements which are inappropriate. Even if we agree to allow multiple ways of recognizing the out-of-place elements--that is, even if we are willing to record certain strings as ambiguously deviant--we still must face the ill-defined problem of determining which portion of a deviant string provides the framework within which the rest can be described as out of place.

For any attempt to deal with this task, we have to distinguish between a deviant string of words taken in the abstract and a deviant or mistaken utterance. We will find for the former that there is simply no possibility of determining in any absolute way its degree of departure from full grammaticality. In the latter case, an account of deviant utterances must take two cases into account: mistakes, as in the speech performances of children, drunkards and foreigners (and the rest of us when we are off our guard), where what is of interest is a comparison between what was intended and what was said; and figurative speech, where what is of interest is the structural type which the speaker wants the hearer to perceive as the framework upon which the hearer's 'construing' abilities can impose some sort of interpretation--hopefully the intended interpretation.

To see what is involved for strings of words considered *in vacuo*, we can take the most favorable case --that of strings which happen to be identical to sentences generated by a grammar which differs in minor ways from the grammar which provides the measure. Suppose, for example, that we wish to say something about the sentences produced by a speaker of a nonstandard dialect of English and suppose that we wish to determine whether it makes sense to talk about the degree of deviation of his sentences from those of the standard dialect.

Taking the single sentence (1), what we need to know first of all is whether it is to be compared with (2) in the standard dialect or with (3).

(1) I seen it.
(2) I have seen it.
(3) I saw it.

Depending on which of these is taken to be the basis of comparison, the sentence is deviant either by virtue of an omission or by virtue of a substitution. If the index we need is something which grades strings of words along the grammaticality dimension, it must be a meaningful question to ask whether the string comes out as more ungrammatical under one of these interpretations than under the other, and it must likewise make sense to ask whether the intuitions of native speakers of the standard dialect can be called on to decide which interpretation is correct. Such inquiry, surely, does not lead to an understanding of where (1) fails with respect to the standard dialect, and we are motivated to look for other kinds of information to tell us this.

Of course, in order to know which comparison is the 'right' one, we need to know whether the rules of the dialect from which we have taken our sample allow the perfect auxiliary 'have' to be contracted to zero (where the standard dialect requires retention of the final fricative), or whether these rules specify 'seen' as the preterite form of 'see'. In case the source dialect has nothing corresponding to the standard dialect contrast between (2) and (3), our problem is more serious still: are we to say that the dialect has only the perfect form, with the auxiliary deleted; that it has only the preterite form, realized phonologically as 'seen'; or that, having the two constructions distinct at some level of analysis, the grammar neutralizes them in surface sentences? The answers to these questions involve detailed comparison of the grammatical rules of the separate dialects, but can in no meaningful way, as far as I can tell, be expressed as information about (1) as viewed from the standard dialect.

With (1) we have the simplest possible case, and yet there were these uncertainties. The situation with random word sequences is totally beyond hope. That becomes obvious as soon as we realize that the possibilities available for matching any one of these with a set of grammatical sentences include the operations of order change, insertion, deletion or replacement of elements, and unrestricted combinations of these.

For utterances that are deviant by mistake, the relevant comparison is between the actual utterance and the intended utterance, but in this case, (i) it is not always possible to know what the intended utterance is, and (ii) it does not matter whether the actual occurring utterance is, in the abstract, grammatical or not.

What is needed is some apparatus for pairing any strings of words with any structural description, and providing some index of the degree of fit between the description and the string, the value of this index determined by an operation which relates the lexical information associated with the individual words of the string with the structural description. Such a device is what we find elaborated in Lakoff (1965). By Lakoff's procedure, any string will have an indefinitely large number of grammaticality values according to the infinite number of structural descriptions that can be brought into association with it. For a fully grammatical sentence there will be at least one structural description which it satisfies completely. An ambiguous grammatical sentence will show perfect fit with two or more structural descriptions--one for each of its possible interpretations. Working out the details requires giving different weight to distinct types of 'poor fit'. All such decisions will involve appeals to native-speaker judgments of some sort, but technically the thing seems feasible.

But notice what happens to our understanding of the working of a generative grammar when we adopt Lakoff's device. The syntactic component specifies the set of well-formed structural descriptions. The dictionary component associates with each lexical item a set of syntactic, semantic, and phonological properties, the syntactic properties understood as including information about insertability into deep-structure configurations and sensitivity to grammatical rules. The relative grammaticality algorithm automatically assigns a grammaticality index to each ordered pair in which the first element is a sequence of lexical items and the second is a structural description.

Under Lakoff's proposal a generative grammar can do what I think Chomsky suggested a generative grammar ought to do, i.e., serve as a grammaticality-index assigning mechanism. But the whole thing depends crucially on having correct information about the lexical items of the language. How are we to discover, our ordinary working grammarian asks, what are the correct lexical properties of the words and morphemes of a language? Can it be, he frets, that the difficulties of knowing correctly the grammar and semantics of lexical items are of the same order of magnitude as those of determining the grammaticality of sentences?

These worries of his are, I think, justified. Presumably, we are to determine the grammatical properties of lexical items by comparing deviant with non-deviant uses of them. We know that 'resemble' is unpassivizable, for example, because speakers of English tell us that while (4) is grammatical, (5) is not.

(4) John resembles a horse.
(5) A horse is resembled by John.

But, in fact, there are some speakers of English who tell us that the passive sentence is not ungrammatical. That means that when we observe a seemingly deviant use of a lexical item we must ask whether this usage constitutes a departure from conventions provided by that speaker's language, whether the speaker's language differs in relevant ways from the language we have been considering, or whether his judgments on grammaticality are sometimes inaccurate. In other words, we must be able to ask whether the speaker regularly uses the word in ways of which the observed usage is an instance, or whether in this situation he made a mistake.

Two examples will demonstrate the difficulty in knowing what the facts are. The first is an elementary case of figurative speech. While it is certainly possible to come up with clear cases, it is frequently in practice impossible to know, even in one's own speech, whether a word has been used figuratively, in the creative sense, or whether it is simply polysemous in the needed way. The use of the word 'bitch' in referring to an unpleasant adult female human was clearly figurative in its first instance, but when we find people who hesitate to use the word when speaking of a female dog, it is apparent that for them the insulting sense of the word involves no appeal to their creative abilities. A description of this state of affairs in terms of the marking of deviance would run like this: somebody whose lexicon contains only the literal interpretation of the noun but who is observed to use it nevertheless when referring to human beings has made a creative extension of the scope of the word that is accounted for by reference to the knowledge that participants in our civilization use attributions to human beings of non-human animal properties for pejoration; somebody who does not use the word when referring to female dogs lacks the original sense and has a lexical entry for 'bitch' with the pejorative sense built in rather than acquired by a construal principle.

Unfortunately, an empirically indistinguishable account is found in the claim that some speakers have two descriptions of the word, others only one. On this interpretation, the acquisition of the non-literal sense is an event in the history of the language. I know of no reasonable proposals for evaluating these alternative accounts.

For a second example, I turn to the fact that some speakers of English do not use 'convince' in the same ways they use 'persuade'. They allow themselves to say (6) but not (7).

(6) We persuaded him to come.
(7) We convinced him to come.

Suppose, knowing that, we hear our informant say (7). We may say that his internal grammar makes the distinction just mentioned, but that he has generalized the infinitive complement construction to the verb 'convince' this one time; or that he is in the process of acquiring the more generalized rule; or that he was imitating speakers of a lesser dialect; or that he mistakenly produced this utterance by choosing the word 'convince' when he intended 'persuade'; or, of course, we might simply say that in his lexicon 'convince' and 'persuade' are given, apart from their phonology, identical descriptions.

There are, then, uncertainties about the proper way of interpreting apparently different uses of lexical items and uncertainties about the accessibility of correct lexical information in general. Appeals to introspection, the compilation of questionnaire results, and claims about idiolectal variation seem not always to point to the truth. Grammatical theory needs instead to consider deviance marking as a precise formal problem, and this it can do by applying to lexical descriptions something akin to Lakoff's proposal for computing relative grammaticality. The lexicon is a device which characterizes well-formed lexical entries but fails to associate phonological material (i.e., 'lexical items') with lexical descriptions. Grammatical theory can now be thought of as providing a way of registering the degree of grammaticality of word strings with respect to structural descriptions *if the lexical descriptions of the words are known*. This is accomplished by associating any sequence of clusters of lexical features--minus the phonological content--with any structural description. The grammar is able to assign indices of relative grammaticality, but only to ordered pairs of lexical description sequences and structural descriptions. The grammar says, in effect: if you can find strings of words that have such-and-such properties, then I can tell you exactly how well they fit any structural description.

If this is what a generative grammar is to do, it has managed to get as far as possible from its initial goal of specifying the well-formed sequences of words. The fact is, of course, that when we took this step we completely lost the attention and interest of our ordinary working grammarian. He wants to know just what these deviance markings are for, and he has serious doubts about whether the speaker's intuitive judgments on grammatically deviant sentences can be accounted for in general in terms of misordering errors and category substitutions of the sort he sees this device capable of detecting. Our grammarian knows first of all that the construal principles for a great many instances of metaphor involve understandings about objects and events rather than properties of the linguistic elements which give expression to these objects and events. More than that, he can think of many cases of what he insists on considering deviant uses of language but which cannot be described by any of the grammar-bound plans for characterizing that have been proposed.

I have in mind a situation like the following. Journalists these days have been made conscious of the jeopardy to justice (or at least the danger of a libel suit) that results from public assignment of guilt to their fellow citizens. They have been instructed to heed certain rules of thumb that are supposed to keep them out of trouble, and among these, I assume, are the following: "Never say of a person who committed a crime that he did it, only that he allegedly did it." "Never call the person who committed the crime the culprit, or the murderer, or the burglar, until after the trial; call him instead the suspect."

As a result of sincere obedience to these injunctions, journalists (perhaps most noticeably in Columbus, Ohio) have acquired odd uses of the adverb 'allegedly' and the noun 'suspect'. Recently I heard on the evening television news in Columbus:

(8) Six members of the Students for a Democratic Society were charged with allegedly distributing inflammatory literature.

(I am assuming, incidentally, that they were charged with actually distributing inflammatory literature; if they were only charged with allegedly doing this, then they were surely guilty, and my point is lost.) In a report on the burglary of a milk store in my city, the local evening newspaper reported that

(9) The police have no clues as to the identity of the suspect.

There was of course no suspect: they had no clues on the identity of the burglar.

These are assuredly deviant uses of the words in question, and I believe they would be recognized as such by their authors if they had had time to edit what they had written. But it seems to me that a correct description of the nature of the deviance is not the sort of thing that can be provided by a generative grammar rigged to assign grammaticality indices. I may be wrong, but I find it difficult to imagine how such an algorithm could successfully mark the two sentences I came across as being more acceptable in journalese than such technically equally odd sentences as (10) or (11).

(10) He wanted the children to allegedly rob the flower-girl.
(11) I hope no suspect burns our house down while we're on vacation.

The deviant uses I have been discussing simply do not involve category errors of familiar kinds.

Uncertainties about the ways in which lexical items figure in the operation of a deviance-marking apparatus brings one face to face with the question of analogy in speech behavior. Although I have agreed with and once contributed to the body of unkind words people have directed toward a little book called State of the Art (Hockett, 1968), I find myself convinced that in the description of changes in the lexicon, the appeal to changes in the content of grammatical rules faces a number of serious difficulties. Consider the recent popularity of event nouns used in the context of social protest in which the first element is a verb and the second element is the preposition 'in', as in 'sit-in', 'love-in', etc. I believe I am correct in my understanding that 'sit-in' was the first of these. The ordinary working grammarian in me wonders how we are to describe what happened when 'sit-in' became a part of the English lexicon. Were there changes in the derivational rules of the language? Was it registered as an unanalyzed lexical item? Or what?

If 'sit-in' entered the language as an unanalyzed lexical item, then it had no influence on the rules, since only generative rules assign structural descriptions. If the word did have an analysis, then there either must be some supplementary apparatus for assigning structure to lexical items, or it must be taken as being generated by a possibly newly created generative rule.

Suppose we take this last position, since it is the only one that is intelligible within the framework of generative grammar. What is the nature of this newly created rule? If the rule is stated as one which takes any verb, shall we say that 'sit' was marked, for a while, as the only verb to which it could apply? Shall be say that the scope of the rule was perfectly general, and merely observe as a fact about the history of usage that nobody bothered to use it for anything but the verb 'sit' for the first few months after the introduction of the rule? (If the answer to this second question is yes, then we must understand the occurrence of the later words in the way that we understand the constructibility of novel sentences.)

But if the original rule was an exceptional one, applying only to 'sit', then what are we to say about such later additions as 'wade-in', 'pray-in' and 'strip-in'? Are we to say that at the later stage the rule became generalized so as to include any verb, or any of a certain type of verb, or are we to say that the grammar became more complicated by virtue of having the relevant exception features added to the verbs 'wade', 'pray', 'love' and the rest? If we accept that the rule was originally general enough to include any verb, in some strict sense of 'verb', was it in fact general enough to include the later hippy creation 'be-in'? If not, with the extension to 'be' are we to say that the rule was further generalized or that it was made more specific so as to include 'be'?

These are all, quite obviously, senseless questions. It would never occur to anyone today to line up all these alternatives and to worry seriously about which is to be preferred, if only because we remember how silly certain older works seem in which we are taught five alternative analyses of the word 'took'. We have here one of those cases where we might indeed agree to say, with Hockett, that somebody made up a word, the word caught on, other people apprehended a pattern and made up some new words on the same pattern. A reconstruction of this history in the form of a sequence of changes in the systems of generative rules would strike the ordinary working grammarian as nothing more than allegiance to a ritual form. However we eventually manage to deal with descriptive problems of this sort, it is at least very clear that in none of this inquiry would it have been of any help to have available to us a metric of relative grammaticality.

5. I have said that it is difficult to see how a generative grammar can be required to demarcate all and only the grammatical sentences of a language in view of some rather serious questions about the empirical determinability of that set; and I have said that it is impossible to imagine any way in which a generative grammar can assign grammaticality indices to deviant sentences. I turn now to a brief consideration of the ways in which a grammar assigns structural descriptions to the sentences which it generates.

The theory of transformational grammar makes available for structural descriptions of sentences (i) the categories of the base rules, (ii) the domination relations that are defined initially by the rules of the base and are adjusted by the transformations, (iii) the left-to-right sequence of elements, (iv) information about permitted co-occurrences in particular structures and (v) information found in the lexicon regarding (a) insertability into deep-structure configurations, (b) sensitivity to grammatical rules, and (c) the semantic structure of lexical items. A grammar is judged as adequate in one important respect if it describes sentences in ways which match certain sorts of intuitive judgments on the part of native speakers, if it captures certain aspects of their knowledge about the sentences.

One specific descriptive problem, ordinarily taken to be the easiest, is that of knowing whether a grammar gives the correct constituent-structure analysis to the surface sentence. Considering the variety of ways in which complex verbal expressions in English get parsed, I am ready to assume that native-speaker intuitions about constituent structure are among the least important criteria for judging the adequacy of proposed descriptions.

But it is also likely that there are a great many facts about the grammatical interpretation of sentences which the devices of categories and sequence and domination fail to capture altogether, yet which must be a part of the generative grammarian's added burden if the goal of achieving descriptive adequacy is to be seriously sought after. I have in mind a number of descriptive problems connected with the treatment of focus, topicalization, reference, deep structure cases, presuppositions, and illocutionary act potential. The brute force method of incorporating all these matters into the theory is by letting assertions about them find their place in proposed underlying structures for sentences. The people called generative semanticists have been accumulating reasons according to which the underlying linguistic structure of the sentence

(12) Did I give you the other book?

will ultimately have to be something which, when rendered into English, would sound like this:

> (13) There is a set of books that both you and I know about and the cardinality of that set is some number n and you and I have just had in mind a subset containing n-1 of those books and I am now calling your attention to the remaining nth book. There was a time when I had that book in my possession and I am now asking you to tell me whether I did anything in the past which would count as causing that book to be in your possession.

The speech act function of the sentence is made explicit in the part about the speaker's requesting an answer from the hearer; the presuppositions are captured in the clauses preceding the operative clause; the category of definiteness is reconstructed as a set of assumptions about what the speaker believes the hearer to be 'having in mind'; and so on.

When the ordinary working grammarian sees such demonstrations, he is properly overwhelmed, but he has trouble believing that the principles by which these maximally abstract representations are to be mapped into the sentences of his language are principles that today's grammarians are equipped to discover. He feels, in fact, that he finds himself in the age of what we might call the New Taxonomy, an era of a new and exuberant cataloguing of the enormous range of facts that linguists need eventually to find theories to deal with. The attempt to capture fully the native speaker's intuitions about the structure and content of his sentences has led to observations which make it extremely difficult to believe in the simple and comforting things we believed in, about grammatical theory, just a few years ago.

6. I see in much recent work a shift of interest away from the properties of an apparatus needed solely for generating the proper set of sentences, toward the mechanisms which speakers of a language can be shown to have, on the basis of any evidence within reach, which account for their ability to do what they do when they communicate with each other using their language. This switch of emphasis to the system itself, and away from the in-or-out judgments associated with the strict notion of generative grammar, makes it possible to ask new kinds of questions. Let me give an example of what I mean.

When grammar-construction is seen as a purely formal task, one of the desiderata of a grammar must be its completeness. In evaluating a grammar which is to generate all and only the sentences of a language, we cannot tolerate a situation in which symbols are introduced at one point and never interpreted or operated on by later rules. It is possible, I want to suggest, that a grammar which exhibits the workings of a natural language cannot meet such a requirement.

It may be that an earlier portion of a grammar allows the introduction of a structure even though the remaining rules of the grammar fail to assign it an acceptable surface form. For types of phenomena that have concerned Perlmutter (in Perlmutter 1968), such a failure is to be accounted for in terms of surface-structure constraints. Surface-structure constraints, however, make up a fairly clearly-defined segment of the grammar itself, and their justification is based on their contribution to the task of isolating grammatical from ungrammatical strings. The issue I am about to bring up is different.

In general, tag questions in English are constructed by adding to any assertive sentence an interrogative piece which contains as subject a pronoun which matches the surface subject of the main sentence, and a pro-verb-phrase which corresponds to the predicate of the main sentence and which is negative in case the main sentence is affirmative, and vice versa. What we need to be able to say about English is that a tag question formative can be chosen with any assertive sentence but the rules for constructing tag questions out of such combinations fail to cover all cases.

People have trouble with tag questions after such sentences as

(14) Somebody's out there.
(15) Somebody tried to get in.
(16) I'm competent to do that.
(17) One of us could go.

The rule for forming the tag question requires the selection of an appropriate pronoun. 'Somebody' is human and singular and unmarked for gender. 'It' is non-human, 'he' and 'she' are marked for gender, and 'they' is plural. There is no pronoun which matches 'somebody'. From the paraphrasability of (14) with (18), many people say (19), but others end up with (20) or (21), and still others give up.

(18) There's somebody out there.
(19) Somebody's out there, isn't there?
(20) Isn't he?
(21) Aren't they?

For a sentence like (15), some people say (22), and others give up; I have heard myself say (23). For (16) some people accept (24), a great many allow themselves to say (25), but many others simply do not know what to say. For (17), the best thing is to make a joke out of it, as in (26). Our grammar sometimes fails us.

(22) Didn't they?
(23) Didn't there?
(24) Aren't I
(25) Ain't I?
(26) One of us could go, couldn't you?

Observations like these are certainly familiar, and for illustrating my point I could just as well have considered the rules for subject-verb agreement and their failure to yield grammatical sentences corresponding to (27) and (28).

(27) Either he or I is? always on duty.
(28) Either he or I am? always on duty.

The recognition of problems of this sort is the recognition of what people try to say, how their grammars fail them, and how eventually they invent a new form, they go ahead and say something they feel is ungrammatical, or they give up. To account for such situations we must allow grammars to be 'incomplete' in just the right ways, that is, for just those situations in which the creative part of a grammar sets up something which the interpretive part cannot cope with.[2]

7. The ordinary working grammarian learns what he can about the grammatical processes which are available to the producers of sentences, and he uses what he knows of these processes for describing these sentences. He welcomes Chomsky's discussions of the non-accessibility of correct grammaticality judgments, because without the Clear Cases Principle to guide him, he knows of no way to bring to his task of writing a grammar the evidence of grammaticality judgments. He wants to know what sorts of things can go wrong in the production of an utterance, and what kinds of freedom creative users of language have for constructing sentences or near-sentences in their language. He does not want to be responsible for a relative grammaticality ranking of utterances or utterance/description pairs.

He will be glad if he can be reassured that his success as a grammarian will not be measured on the basis of his ability to demonstrate that his grammar does everything that generative grammars have been said to have to do. I believe he deserves such reassurance.

Knowing what he does not have to do will not give him reliable insights into what he does have to do, unfortunately, but that is because the ordinary working grammarian I have in mind is exactly as confused as I am about that. If he is a practitioner of the New Taxonomy, he is having a good time. It is possible to remain happy, for a while, without well-defined goals.

[2] It should be pointed out, incidentally, that the discovery of this sort of operative failure in a grammar offers no comfort to those persistent spokesmen for the inherent vagueness of grammars. Grammars may indeed have areas of unimprovable vagueness, but the facts about English that I have been discussing can be made totally explicit. What gives the native speaker the impression of vagueness is his uncertainty about knowing what to do when he wants to say something which his grammar--in ways unknown to him--fails to allow him to say.

References:

Chomsky, Noam (1957) Syntactic Structures, Mouton.

Chomsky, Noam (1965) Aspects of the Theory of Syntax, M.I.T. Press.

Chomsky, Noam (1966) "Topics in the theory of generative grammar," pp. 1-58 of Thomas A. Sebeok, ed., Current Trends in Linguistics Vol. III: Theoretical Foundations, Mouton.

Chomsky, Noam and George A. Miller (1963) "Introduction to the formal analysis of natural languages," pp. 269-321 of R. D. Luce, et al., Handbook of Mathematical Psychology, Vol. II, Wiley Press.

Elliott, Dale, Stanley Legum and Sandra Annear Thompson (1969) "Syntactic variation as linguistic data," pp. 52-59 of Robert Binnick et al., eds., Papers from the Fifth Regional Meeting of the Chicago Linguistic Society, University of Chicago.

Hockett, Charles F. (1968) The State of the Art, Mouton.

Katz, Jerrold J. (1966) The Philosophy of Language, Harper and Row.

Lakoff, George (1965) On the Nature of Syntactic Irregularity, The Computation Laboratory of Harvard University Mathematical Linguistics and Automatic Translation, Report No. NSF-16. Cambridge, Mass.

Perlmutter, David (1968) Deep and Surface Structure Constraints in Syntax, M.I.T. dissertation.

Ross, John R. (1967) Constraints on Variables in Syntax, M.I.T. dissertation.

Thorne, James Peter (1964) "Grammar and Machines," pp. 293-306 of R. C. Oldfield and J. C. Marshall, eds., Language, Penguin.

# Relative Clause Structures and Constraints on Types of Complex Sentences*

Sandra Annear Thompson

*Sponsored in part by the National Science Foundation through Grant GN 534.1 from the Office of Science Information Service to the Computer and Information Science Research Center, The Ohio State University.

# Relative Clause Structures and Constraints on Types of Complex Sentences*

Sandra A. Thompson

The study of relative clause sentences shows quite clearly the limitations of our present grasp of the relation between syntax and semantics. In trying to understand what relative clause sentences are, I have been struck by two facts. First, there are certain very clear structural properties of relative clause sentences which distinguish them from sentences containing. for example, sentential subjects or objects. These seem to me to be strictly syntactic properties. Second, there are certain very subtle meaning differences between definite and indefinite noun phrases containing relative clauses; clearly a semantic fact. Let us briefly characterize these two facts.

Describing the semantics of definiteness seems in part to involve, as outlined by Karttunen (1969), several disjunctive statements of what the speaker presupposes the hearer knows about the entity named by the definite noun phrase. But this would be just a beginning, for we find the definite determiner used with superlatives, and we find certain instances of the conditions for its use apparently being satisfied, and yet it is not used. For example, I am puzzled as to why we generally say

(1) Here is a cookie.

or

(2) Have a cookie.

instead of

(3) Here is the cookie.

and

*This paper and the other paper of mine appearing in this volume constitute the revision of my Ph. D. dissertation, received at Ohio State University, December, 1969. I am indebted to the following people for valuable discussion on various aspects of this work: George Bedell, Dale Elliott, Jim Heringer, Shuan-fan Huang, Andreas Koutsoudas, Steve Krashen, Stan Legum, Peter Menzel, Barbara Partee, Jerry Sanders, Robert Stockwell, McMillan Thompson, Arnold Zwicky, and especially to my teachers Terry Langendoen, Ilse Lehiste, and Charles Fillmore for their investment of time and interest in my work. Of course, none of these people necessarily agrees with everything I say here.

(4) Have the cookie.

under conditions that normally give rise to the definite determiner, namely in case the referent is in sight of both speaker and hearer, as in

(5) Look at the dog.
(6) Please lock the door.

I look forward to a discussion of a model of linguistic description which allows for an account of such semantic areas as definiteness; I see no proposal of such a model even on the horizon at present.

The structural facts I have been referring to include the following: Relative clause sentences appear to be instances of the syntactic device known as embedding. However, when we compare them to sentences containing embedded subjects and objects as in

(7) His speaking so eloquently impresses me.
(8) I like his speaking so eloquently.

we find that the embedded portions of (7) and (8) play an obligatory role with respect to the main verb, the role which Fillmore (1968) has called the objective case. In addition, the verb governs both the occurrence of the clause in such sentences and the type of clause which can occur. These facts are not true of relative clause sentences. The relative clause plays no role, obligatory or otherwise, with respect to the main verb. No verb, then, is ever marked for taking a relative clause. Structurally speaking, it is superfluous. The relative clause sentence amounts to two independent propositions. These facts can be accounted for if only sentences with embedded subjects and objects are considered to be instances of underlying embedding, and relative clause sentences are taken as instances of underlying conjunction.

The facts that lead me to this conclusion seem to be quite independent of whether the head noun in the relative clause sentence is definite or indefinite. There is no structural motivation for assuming a different underlying representation for these two sentences:

(9) The pitcher that I gave to Harry last year is on Jane's table now.
(10) A pitcher that I gave to Harry last year is on Jane's table now.

Definiteness is simply not relevant for specifying these syntactic facts about relative clause sentences.

The approach which appears to me to be incorrect, that there is no underlying autonomous level of syntactic representation, has been argued for by Lakoff (to appear), Lakoff and Ross (1967), McCawley (1967, 1968a, 1968b, in press). I object to this position on two grounds.

First, it suggests that there is, if we can just get deep enough, one abstract representation (referred to by the proponents as "the semantic representation") for a given sentence. But there is no reason to believe that one deep structure cannot underlie more than one sentence, even where the sentences are not necessarily synonymous.

My second point of criticism is that it seems quite possible, and I think conceptually valuable, to define a level of representation exhibiting the relationships among the items in one sentence as well as the relationships among the simplex components of a complex structure. The question of whether these relationships are semantic or syntactic simply does not need to arise. The fact is that these relations among the structural elements of a sentence can be represented formally in terms of structures which underlie surface structures exactly according to the kinds of arguments for underlying structures which can be found in Postal (1964) and Chomsky (1965). Moreover, it is only information of this relational type which has been shown so far to justify any formalism for underlying representations. In other words, we can conceive of a theory of sentence structure in terms of a model in which superficial syntactic structures are related to more abstract syntactico-semantic structures. These are the deepest representations which we are able to discover by the use of linguistic evidence; they represent the relationships among the basic pieces of a sentence. This view will be elaborated in more detail below.

In other words, as long as we insist that sentences (9) and (10) must have different deepest underlying representations because they have recognizably different meanings, the structural relationship between relative clause sentences and conjunctions will be virtually impossible to discern. Indeed this relationship has not been observed, and I think this is because it is very difficult to represent one pair of conjuncts as distinct from another in such a way as to capture the meaning difference between (9) and (10). I, for one, cannot construct such conjunction sets.

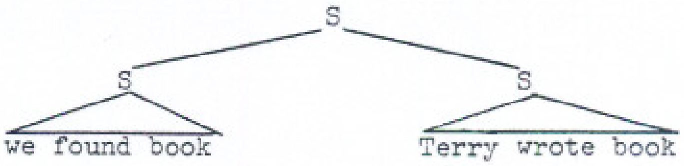
In fact, I think this is totally incorrect. My understanding of the structure of relative clause sentences and sentence complexity depends upon the assumption that the deepest structure of a sentence does not exhibit the full range of information about its meaning. I will suggest that what underlies a sentence is not a single representation revealing everything that that sentence "means," but instead a "basic elemental structure" (BES) for that sentence plus a set of parameters which can be associated with this BES and which play a role in determining the transformations which it can undergo and what its surface structure will be. The BES will specify the meanings of "content words" and the relationships among them.

To see how I suggest that syntactic structures and semantic "parameters" might interact, let us take an example. Underlying the sentence

(11) We found the book that Terry wrote.

would be the BES:

(12)



The way this structure comes to the surface is determined partially by such factors as definiteness. If the speaker presupposes[1] that the hearer knows neither of the facts expressed in the conjunction, the surface form of the sentence may be the conjunction

(13) Terry wrote a book and we found it.

or either of the following relative clause sentences, with an indefinite head noun:

(14) We found a book that Terry wrote.
(15) Terry wrote a book that we found.

If the speaker presupposes that the hearer knows about the fact expressed in the first conjunct, the corresponding relative clause sentence will have that conjunct as the relative clause, and the head noun must be definite:

(16) Terry wrote the book that we found.

and if the speaker presupposes that the hearer knows about the information of the second conjunct, the surface sentence will be

(17) We found the book that Terry wrote.

The implications of this analysis as well as well as the relationship between non-restrictive relative clause sentences and conjunctions are discussed more fully in my paper, "The deep structure of relative clauses."

This account is the only one I know of which attempts to characterize the fact that, with the definite determiner, the relative clause expresses information which the speaker presupposes to be known by the hearer.

It might be suggested that, instead of having certain semantic parameters affecting the derivation of a sentence at various underlying levels, I could as easily have semantic interpretation rules assign appropriate readings to the surface structure of the sentence

[1] "Presupposes x" is here used in the fairly well-accepted sense of "believes x," where x represents a set of conditions which must be met for a given sentence to be uttered in good faith, and which are independent of the illocutionary force of that sentence.

in question. Although there is much that is similar in these two positions, evidence such as the following, which was provided by Stephen Krashen (personal communication), forces me to prefer the former. The presuppositional facts which hold for relative clause sentences hold for sentences with pre-nominal adjectives as well. The sentence

(18) Janice wore the outfit that is black.

is used by a speaker who presupposes that the hearer knows about the outfit that is black. The sentence

(19) Janice wore the black outfit.

carries precisely the same presuppositions. A semantic interpretation rule providing this information which operated on surface structures would either have to be stated twice, once for each surface structure, or it would have to contain a note to the effect that what is semantically true of relative clause sentences is true of prenominal adjective sentences. Since neither of these alternatives is defensible, I prefer to view the interaction between syntax and semantics as I have outlined above.

In certain respects my view of the relationship between syntax and semantics is reminiscent of that put forth in Chapter 9 of Chomsky (1957):

> It seems clear that undeniable, though only imperfect correspondences hold between formal and semantic features in language. The fact that the correspondences are so inexact suggests that meaning will be relatively useless as a basis for grammatical description. Careful analysis of each proposal for reliance on meaning confirms this and shows, in fact, that important insights and generalizations about linguistic structure may be missed if vague semantic clues are followed too closely. For example, we have seen that the active-passive relation is just one instance of a very general and fundamental aspect of formal linguistic structure. The similarity between active-passive, negation, declarative-interrogative, and other transformational relations would not have come to light if the active-passive relation had been investigated exclusively in terms of such notions as synonymity. (p. 10)

To digress briefly, my position is also reminiscent of the position taken by the opponents of Bertrand Russell's "Theory of Descriptions." [According to Russell (1905, 1919), definite noun phrases which are complex "refer to" entities in a very special way.] Such noun phrases cannot refer to entities in the ordinary sense, he argues, because it is quite possible to reformulate a proposition containing the noun phrase in question without mentioning it at all. So, the sentence

(20) The author of Waverley is Scotch (sic).

is actually a conjunction of three propositions:

(21) (a) At least one person wrote *Waverley*.
(b) At most one person wrote *Waverley*.
(c) Whoever wrote *Waverley* is Scotch.
Russell (1919, p. 177)

A sentence like (20), then, does not have to be viewed as meaningless even if there is no author of *Waverley*; (20) could be asserted to be false because (21a) was false. Similarly, the sentence

(22) The King of France is bald.

can be shown to be false because one of its underlying propositions is false.

This view was criticized, I think correctly, by both Geach and Strawson. According to Geach (1950), when the definite noun phrase, "the king of France" is the logical subject of a sentence, an affirmative answer is presupposed to the question

(23) Does the King of France exist?

Since the answer to (23) is negative, the use of the phrase as a logical subject is out of place, and the question of the truth of a sentence in which the phrase is a logical subject does not arise.

Strawson (1950) objects in a similar vein: it is false to say that a sentence such as (20) contains the proposition (21a) without recognizing that (20) is an assertion and (21a) is "implied" (in some sense of that term) by (20).

Noun phrases containing relative clauses are "definite descriptions" for Russell, similar to the subjects of (20) and (22). While the philosophical problem to which Russell, Geach, and Strawson were addressing themselves is not of central concern here, the conclusion of the latter two thinkers parallels that which I have reached. In fact, I am going one step further: in addition to claiming that the user of the phrase

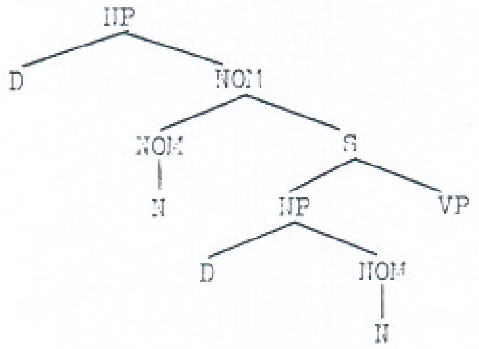
(24) the girl who speaks Basque

presupposes the *existence* of such a girl, I am suggesting that he also presupposes that her existence is known to the hearer.

A. "Stacked" Relative Clauses.

Most accounts of relative clause structures that I know of simply *assume* an underlying embedded structure, a natural assumption to make on the basis of their surface embedded form. The notable exception is the work of the UCLA English Syntax Project (Stockwell et al. (1968)), in which an attempt is made to *justify* the underlying embedded structure. It is argued there that the correct analysis of relative clauses requires configurations of the form

(25)



Stockwell et al.(1968), "Relativization"

The argument for this representation turns on the claim that NP's containing more than one relative clause can be interpreted in such a way that each successive relative clause from right to left "modifies" or "restricts" the meaning of the head noun plus the preceding relative clauses. Let us take an example (op. cit. p. 23):

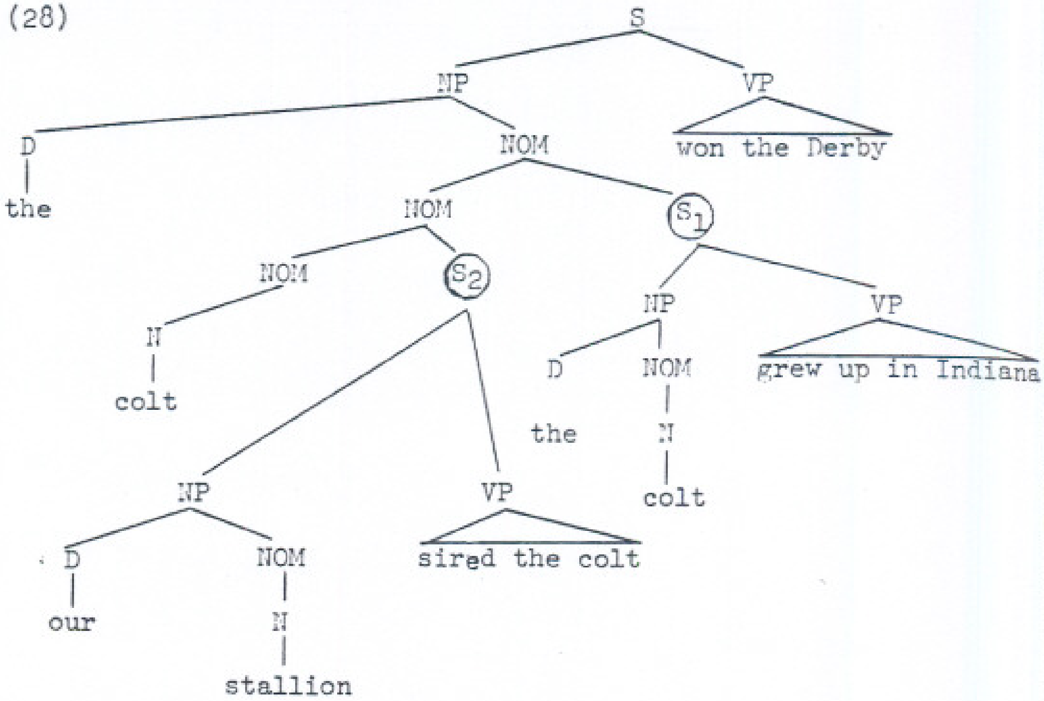
(26) The colt that our stallion sired that grew up in Indiana won the Derby.

One interpretation for (26) is

(27) Out of all the colts sired by our stallion, the one that won the Derby grew up in Indiana.

In this interpretation, the clause that grew up in Indiana is taken as "restricting" the class of objects referred to by the expression the colt that our stallion sired. Let us call an interpretation like that represented by (27) a "stacked interpretation" of a multiple-relation-clause sentence, and let us furthermore call the clause that "modifies" the head noun and the other relative clause in such an interpretation the "higher ranking clause." The structure for (26) according to the UCLA analysis would be:

(28)



If the stacked interpretation of (26) has as its basis a structure like (28), then we have both an argument *for* an embedding analysis of relative clause sentences and an argument *against* a conjoining analysis of such sentences, since a conjoining structure could not directly show that the clause which I have labeled $S_1$ in structure (28) is of higher rank than that which I have labeled $S_2$.

However, as was pointed out to me by C. J. Fillmore, although multiple relative clause constructions can be interpreted this way, there are good reasons for rejecting the proposal that stacked interpretations should be explained in terms of stacked structures.

First, notice that the stacked interpretation is closely correlated with stress. For many speakers, a non-conjoined interpretation is possible *only* if one clause or the other is stressed. Otherwise, the subject of (26) would be interpreted, for these speakers, as referring to a colt that both had our stallion as a parent and grew up in Indiana. For a few speakers, the stacked interpretation is possible with no special stress; for them a normally stressed multiple relative clause sentence is ambiguous between (a) a conjoined interpretation and (b) a stacked interpretation in which the outer relative clause is of higher rank. Incidentally, as pointed out in Stockwell et al (1968), the colt specified by interpretation (a) is the same colt as the one specified by interpertation (b). In other words, a conjoined and stacked interpretation do not differ extensionally.

The crucial point here is that for both of these groups of speakers, the interpretation can be switched so that the inner clause is interpreted as being of higher rank simply by stressing that inner clause.[2] So,

(29) The colt that our stallion sired that grew up in Indiana won the Derby.

cannot be interpreted by either group of speakers except in the following way:

(30) Out of all the colts that grew up in Indiana, the one that was sired by our stallion won the Derby.

What this means is that there is no motivation at all for trying to explain stacked interpretations on the basis of underlying embedding relationships in terms of which the outer relative clause is "higher" than the head noun with the inner relative clause. The fact that a clause is "higher" in a structure like (29) does not seem to correlate with whether it is interpreted as being of higher rank. It might be argued that some device could easily be introduced into the structure which would allow the lower relative clause to be interpreted as the higher ranking one, or that a "clause-scrambling rule" could operate. However, neither of these devices would in any way enhance the proposal that a stacked interpretation is based on a stacked structure, since the purpose of the introduction of either of these devices would be to force a certain interpretation *in spite of* the structure. In other words, the interpretation is independent of the structure, and the theory should reflect that this is the case.

Second, if a stacked relative clause structure could be justified, and did influence interpretation, then we would expect to find that preposed adjectives would carry with them the information as to which clause, in terms of position, they came from. The fact that they do not makes the stacked structure highly suspicious. That is, we can interpret the adjective-preposing rule as operating cyclically, and working from the bottom up, on each cycle inserting the adjective into the NP of the S immediately above it. But notice that no matter whether each adjective is placed *before* the adjective from the next lower S, or *after* it, there is still no correlation between this linear ordering and interpretation of rank. For example, from the noun phrase

(31) the man who has a beard who is bare-footed

[2] The UCLA group has recognized that the stacked interpretation of relative clauses is problematic, but their conclusion is that accepting or not accepting a stacked interpretation is one respect in which dialects can vary, and that to describe this difference might involve postulating different (though both embedded) analyses of relative clause sentences.

we can derive the following noun phrases, by preposing one adjective at a time;

(32) the bearded man who is bare-footed
(33) the bare-footed man who has a beard

or both together;

(34) the bearded bare-footed man

In each case, the interpretation is linked to the stress: the adjective that is stressed is interpreted as modifying the rest of the NP, independent of its position; if neither is stressed, they are interpreted as being conjoined.

Third, if the stacked relative clause structure were a basic, meaning-determining structure, we would expect the interpretation to come as naturally for sentences with indefinite determiners as for those with definite determiners. Actually this is not the case; a stacked interpretation is very difficult to impose on an NP with an indefinite determiner:

(35) A man who had a beard who was wearing a striped shirt was passing out McCarthy buttons.

If we assumed that a stacked interpretation of relative clauses is structurally based, then, given that such an interpretation cannot be imposed on an indefinite sentence, we would have to find a way of blocking such structures in case the definite determiner has not been chosen; or alternatively, we would have to block the choice of the indefinite determiner for this structure. Either of these would seem to be an unfortunate device to introduce, since I know of no other cases in which the choice of the definite determiner depends solely on the <u>structure</u> of the sentence into which it is to be inserted.

What I have shown here indicates that an argument for an embedded analysis of relative clauses which depends on a structural explanation for the stacked interpretation of relative clauses collapses when this explanation is shown to be the wrong one for such an interpretation.

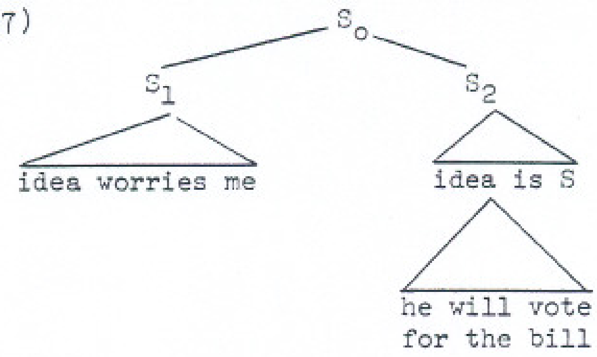
B. Noun Complements.

Relative clauses have been noted to be distinct from embedded clauses in several important respects. Another type of apparently embedded clause which can be shown to behave structurally quite similarly to relative clauses is the noun complement as in

(36) <u>The idea that he will vote for the bill</u> worries us.

as with relative clause sentences, (36) makes two independent prepositions involving the noun <u>idea</u>; the <u>that</u> clause is structurally superfluous. If we postulate (37) as a source for (36)

(37)



we find that precisely the same rules that generate relative clause sentences from conjunctions are used in the derivation of noun-complement sentences as well. From (37), by embedding $S_2$ into $S_1$, we can derive

(38) The idea which is that he will vote for the
bill worries us

Deletion of the WH-form plus BE is obligatory when the BE is followed by a complementizer and a sentence, resulting in sentence (36). Embedding $S_1$ into $S_2$ results in

(39) The idea which worries us is that he will vote
for the bill.

The difference between the two phrases

(40) the idea that we should go to the party
(41) the idea that you mentioned

is clearly that, in the first the *that* is a *complementizer* while in the second it is a relative word, a replacive for another occurrence of *idea*. This correlates with the fact that they do not conjoin:

(42) *the idea that we should go to the party and
that you mentioned

Thus, although both (40) and (41) can be derived from conjunctions (though not the same conjunctions) by the same set of rules, they are not structurally identical. The structural difference also correlates with the fact that relative clause noun phrases, like

(43) the dog that they bought

and noun-complement phrases as

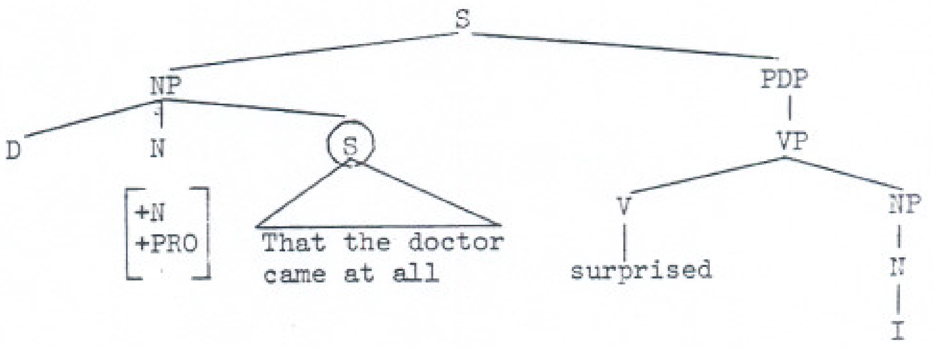
(44) the idea that she's a mother

normally receive different intonation patterns.

One might object that in my analysis the parallelism between the claim that S and I claim that S is lost. It seems quite reasonable, however, to consider that the lexicon shows that such noun-verb pairs are listed with the information that where the verb takes a sentential object, the noun takes a sentential predicate nominal.

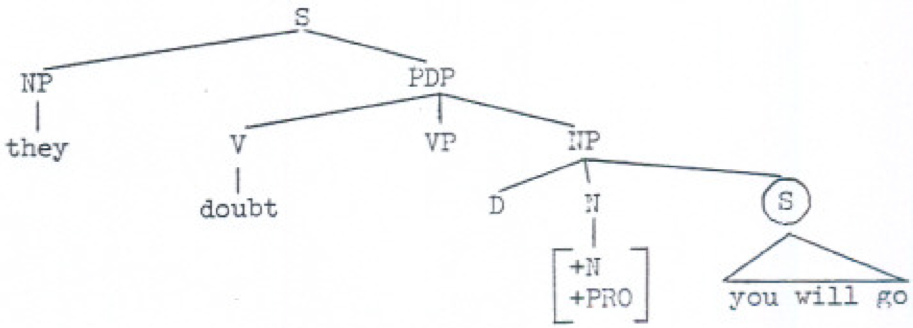
Placing the remarks we have just made concerning the nature of relative clause and noun complement sentences in a broader perspective of sentence complexity, let us consider the variety of sentence types for which embedding analyses have been proposed. The embedded S in each case has been circled.

(45) That the doctor came at all surprised me.



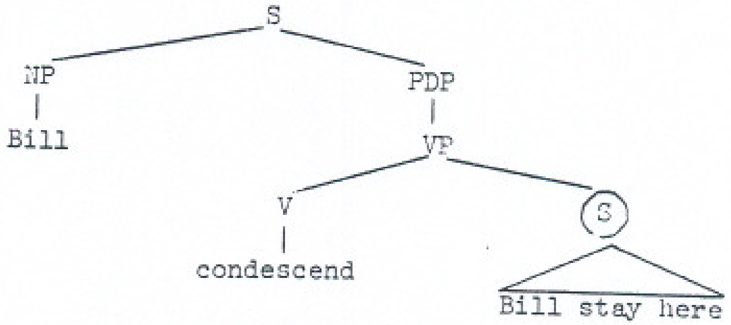
Rosenbaum (1967), p. 12

(46) They doubt that you will go.



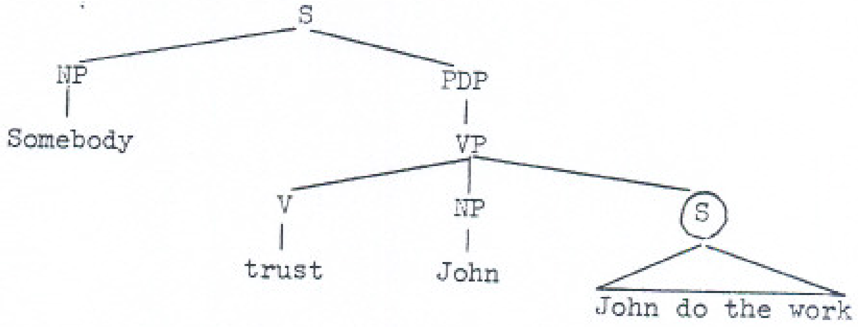
Rosenbaum (1967), p. 34

(47) Bill condescended to stay here.



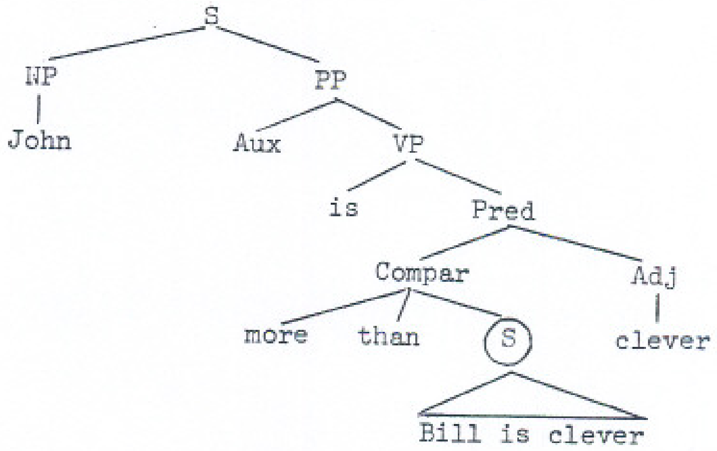
Rosenbaum (1967), p. 94

(48) Somebody trusts John to do the work.



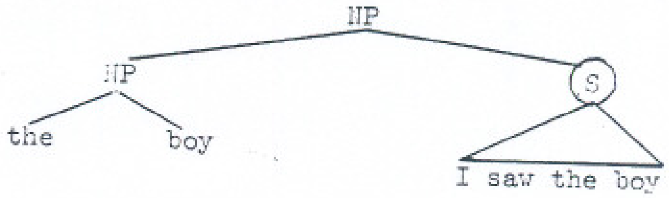
Rosenbaum (1967), p. 9

(49) John is more clever than Bill.



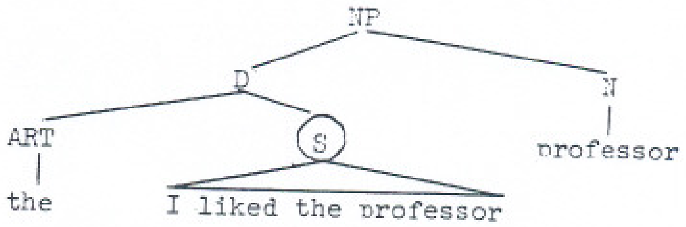
Chomsky (1965), p. 178

(50) the boy I saw



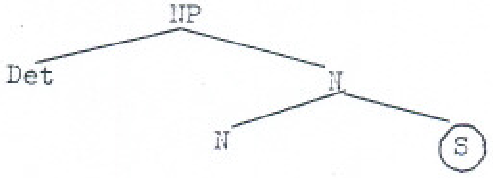
Ross (1967), p. 184

(51) the professor I liked



Stockwell et al. (1968), "Relativization," p. 3

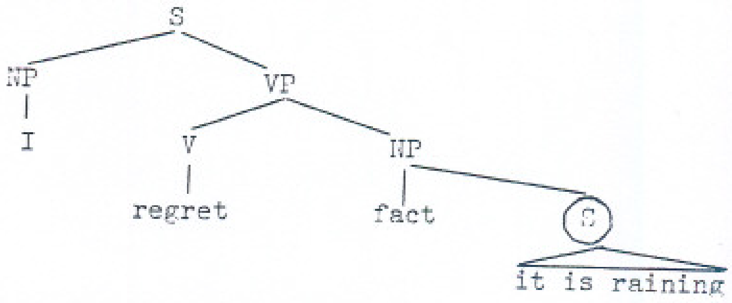
(52)



Dean (1967), p. 34

The previous three structures for relative clause sentences are in competition.

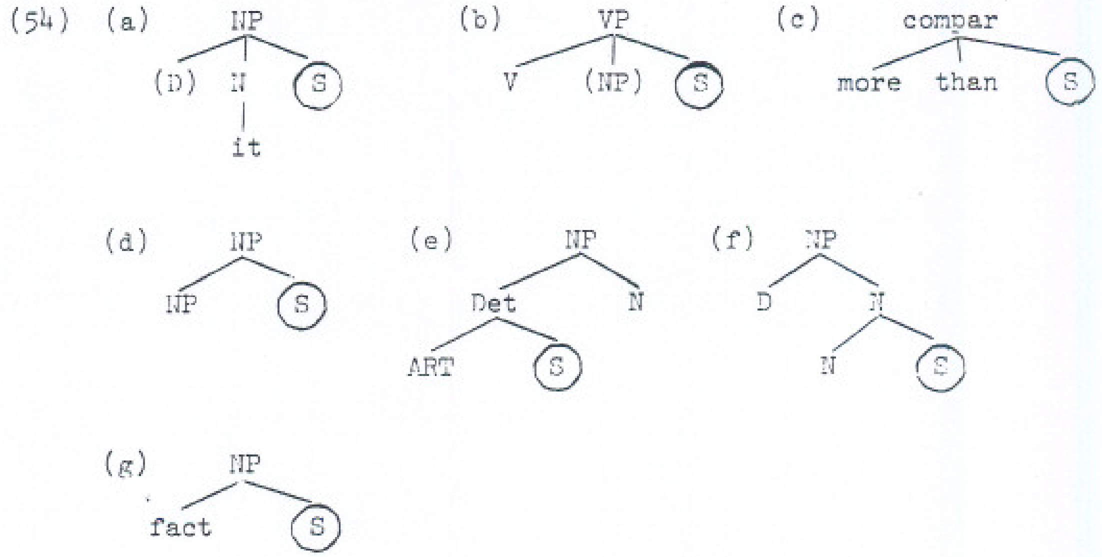
(53) I regret that it is raining.



After Kiparsky (forthcoming)

The following set of configurations summarizes the environments in which it has been proposed that S's be introduced:

(54)



3 The fact that each of the structures in (54) has been recently called into question suggests to me that a very natural constraint can be placed on the introduction of embedded S's. The advantages of eliminating both structures (a) and (b) have been fully dealt with by UCLA (1969), "Nominalization," Bowers (1968), and Wagner (1968); a more adequate analysis than (c) for the comparative is provided by Celce (1970); and I have shown here why structures (d) through (g) fail to be the best representations for relative clauses and noun complements. The only instances of embedded S which remain unquestioned are those in which the S is the unique expansion of an NP which is a subject or an object. Extending Fillmore's suggestion (1968), p. 28 to "limit complement S to the case OBJECTIVE," I would suggest that

(55) All occurrences of non-topmost S's not immediately dominated by S be limited to unique expansions of subject or object NP's.

Let us call (55) the "embedding constraint." Such a constraint limits the power of a grammar in an important way by providing a natural restriction on the ways in which complex sentences may be built out of simple ones, and it provides an account of the differences between embedded and conjoined structures which I have pointed out above.

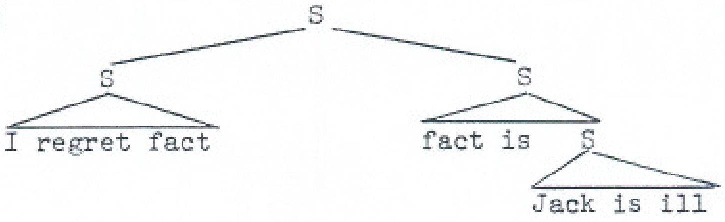
APPENDIX

Some further remarks are in order concerning one member of the class of nouns which take the surface complements, namely fact. I seem to be suggesting that sentences such as:

(56) I regret the fact that Jack is ill.
(57) I regret that Jack is ill.

have different BES's, that (56) would be derived from a conjunction, while (57) would result from embedding *Jack is ill* as the object of *regret*. In fact, I think that the distinction made by the Kiparskys (forthcoming) between factive and non-factive complements is a very important one and I agree with them that sentences like (56) have the same source as sentences like (57). I would propose the following source:

(58)



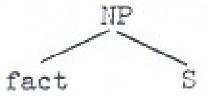
*Regret* is a verb which must take *fact* as its object; application of the rule which deletes *fact* yields:

(59) I regret (Jack is ill).

which more directly underlies (57).

As confirmation of (58), Stephen Krashen has pointed out to me that if a configuration must be involved in stating the presupposition of factivity, as Kiparskys claim, it cannot be:

(60)

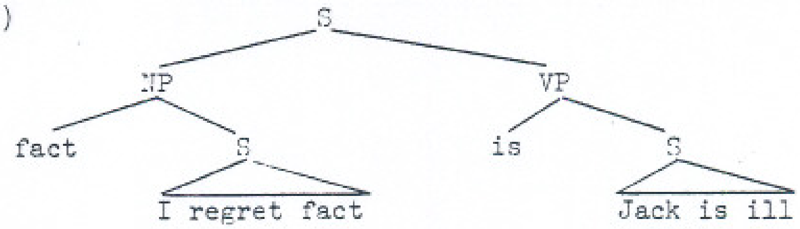


The reason for this is that the truth of the sentence *Jack is ill* is presupposed in:

(61) The fact that I regret is that *Jack is ill*.

exactly as it is in (56). According to my analysis, both (56) and (61) are derived from the same BES, namely (58) (see p. 31 ), so that this semantic fact is accounted for in a natural way. But the Kiparskys would have to have some way besides (60) to account for the presupposition of factivity in (61), since there is no (60) under the VP node in the underlying representation that they would suggest it has:

(62)



In terms of this analysis of factive sentences we are now ready to consider an apparent counterexample to the embedding constraint. A sentence like

(63) That the floor is sticky $\left\{\begin{array}{l}\text{shows}\\ \text{indicates}\\ \text{implies}\\ \text{proves}\end{array}\right\}$ that

Darlene spilled the Koolaid.

seems to contain two occurrences of the OBJECTIVE case, a natural conclusion if complex subjects and objects can arise only from NP's in an OBJECTIVE relationship to the verb. As D. T. Langendoen pointed out to me, however, two facts indicate that the objections raised by this type of sentence are only pseudo-problems for the embedding constraint: (a) the subject clause in (63) is a predicate nominal clause to *fact*; and (b) this noun *fact* is in the INSTRUMENTAL case. To demonstrate point (a), I would cite the obvious paraphrase of (63),

(64) *The fact* that the floor is sticky $\left\{\begin{array}{l}\text{shows}\\ \text{indicates}\\ \text{implies}\\ \text{proves}\end{array}\right\}$

that Darlene spilled the Koolaid.

Furthermore

(65) *(The fact) that the floor is sticky $\left\{\begin{array}{l}\text{shows}\\ \text{indicates}\\ \text{implies}\\ \text{proves}\end{array}\right\}$

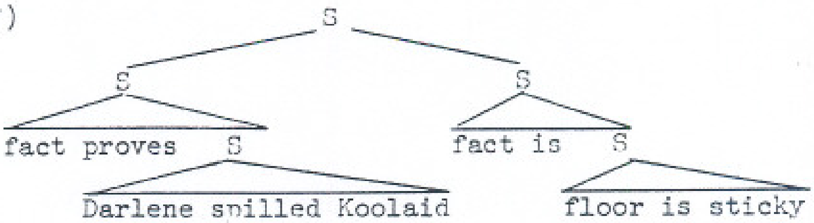
the fact that Darlene spilled the Koolaid.

with *the fact* in both clauses, is ungrammatical, because of a selectional restriction which does not allow *show*, *indicate*, *imply*, *prove*, and the like to occur with factive objects. Sentence (64) now can be shown to be identical in structure to a sentence like (66)

(66) The knife cut the cheese.

That is, a knife can cut cheese only if someone uses it to cut cheese; a fact does not prove something unless someone uses it to prove something. In each case the verb requires an agent, which may be optionally deleted. The *fact*-clause in (64) is thus an underlying instrumental NP, just as *the knife* is in (66). The BES I propose for (64) is:

(67)



References:

Bach, Emmon and Robert T. Harms, eds. (1968). Universals in Linguistic Theory, New York: Holt, Rinehart & Winston.

Bierwisch, Manfred and Karl Heidolph, eds. (forthcoming) Recent Advances in Linguistics, Mouton.

Bowers, Frederick (1968) "English complex sentence formation," Journal of Linguistics 4.1.83.

Celce, Marianne (1970) "The English Comparative Re-examined," Systems Development Corporation publication, Santa Monica, Calif.

Chomsky, Noam (1957) Syntactic Structures, s'Gravenhage, Mouton.

Chomsky, Noam (1965) Aspects of the Theory of Syntax, Cambridge, Mass.: M.I.T. Press.

Darden, Bill, Charles-James Bailey, and Alice Davison, Eds. (1968) Papers from the Fourth Regional Meeting of the Chicago Linguistic Society, University of Chicago.

Dean, Janet (1967) "Determiners and relative clauses," M.I.T, circulated in PEGS.

Fillmore, Charles J. (1968) "The case for case," in Bach and Harms (1968), 1-88.

Fillmore, Charles J. and D. Terence Langendoen, eds. (to appear) Studies in Linguistic Semantics.

Geach, P. T. (1950) "Russell's Theory of Descriptions," Analysis 10.4.

Jacobs, Roderick and Peter Rosenbaum (in press) Readings in English Transformational Grammar, Waltham, Mass.: Blaisdell (Ginn & Col).

Karttunen, Lauri (1969) "Problems of reference in syntax," mimeo, University of Texas.

Kiparsky, Paul and Carol Kiparsky (forthcoming) "Fact," to appear in Bierwisch and Heidolph, eds. (forthcoming).

Lakoff, George (to appear) "On Generative Semantics," in Steinberg and Jakobovits, eds. (to appear).

Lakoff, George and John Ross (1967) "Is Deep Structure Necessary?" ditto, Harvard University, M.I.T.

McCawley, James (1967) "Meaning and the Description of Languages," Kotoba no Ucho, Vol. 2, Nos. 9, 10, 11.

McCawley, James (1968a) "The Role of Semantics in a Grammar," in Bach and Harms (1968).

McCawley, James (1968b) "Lexical Insertion in a Transformational Grammar without Deep Structure," in Darden et al, (1968).

McCawley, James (in press) "Where do Nounphrases Come From?" to appear in Jacobs and Rosenbaum (in press).

Postal, Paul (1964) "Underlying and Superficial Linguistic Structure," Harvard Educational Review 34.246. Reprinted in Reibel and Schane (1969).

Reibel, David and Sanford Schane (1969). Modern Studies in English, Englewood Cliffs, N.J.: Prentice-Hall.

Rosenbaum, Peter (1967) The Grammar of English Predicate Complement Constructions, Cambridge, Mass.: M.I.T. Press.

Ross, John (1967) "Constraints on Variables in Syntax," M.I.T. Ph. D. dissertation.

Russell, Bertrand (1905) "On Denoting," Mind 14.479-493.

Russell, Bertrand (1919) Introduction to Mathematical Philosophy, London, George Allen & Unwin, Ltd.

Steinberg, Danny and Leon Jakobovits, eds. (to appear) Semantics--An Interdisciplinary Reader in Philosophy, Linguistics, Anthropology, and Psychology.

Stockwell, Robert P,, Paul Schachter, and Barbara H. Partee (1968) Integration of Transformational Theories on English Syntax. UCLA.

Strawson, P. F. (1950) "On Referring," Mind

Thompson, Sandra Annear (to appear) "The Underlying Structure of Relative Clauses," In this volume, also in Fillmore, Charles J. and D. Terence Langendoen, eds., (to appear).

Wagner, K. Heinz (1968) "Verb Phrase Complementation: A Criticism," Journal of Linguistics 4.1.89.

# The Deep Structure of Relative Clauses*

Sandra Annear Thompson

*Sponsored in part by the National Science Foundation through Grant GN-534.1 from the Office of Science Information Service to the Computer and Information Science Research Center, The Ohio State University.

# The Deep Structure of Relative Clauses*

Sandra Annear Thompson

A number of general studies in transformational grammar (including Chomsky (1965), Jacobs and Rosenbaum (1967), (1968), Lakoff (1966), Langendoen (1969), Ross (1967)) have assumed that the appropriate underlying representation for a relative clause sentence involves a sentence embedded into a noun phrase. I would like to question this assumption, and to suggest that in fact the appropriate underlying representation for a relative clause sentence is a conjunction.

The argument will be developed in several stages. First, I will suggest some facts which indicate what conjunctions must underlie relative clause sentences. Next, I will show the general process of relative clause formation and some of the implications of my analysis.

Finally, I will indicate in what respects the derivation of sentences containing non-restrictive relative clauses is similar to that of sentences with restrictive relative clauses.

I. Indications that a conjunction source for relative clause sentences is correct.

(a) To my knowledge, no arguments defending an embedding analysis against a conjunction analysis for relative clause sentences have ever been presented either in the literature or informally.

(b) There is virtually no agreement among those who assume that relative clauses are underlyingly embedded as to what configuration of nodes is appropriate to represent the relationship between the two sentences. UCLA (1969) presents a summary of the various approaches which have been taken and the arguments given to support each.

(c) There is a significant but generally overlooked set of structural distinctions between relative clause sentences and those complex sentences which are clearly realizations of structures containing embedded sentences, namely those containing sentential subjects or objects, such as:

(1) That Frieda likes to cook is obvious to me.
(2) I think that Frieda likes to cook.

For sentences (1) and (2), an embedding analysis is well-motivated since the contained sentence is required as an obligatory argument of the verb; it plays a role with respect to the verb which Fillmore (1968) has called the <u>objective</u> role and without which the verb cannot stand. Furthermore, the verb governs both the occurrence

*To appear in D. T. Langendoen and C. J. Fillmore (eds.), <u>Studies in Linguistic Semantics</u>, Holt, Rinehart and Winston.

of clause and the type of clause which can occur. These conditions do not hold for relative clause sentences. A relative clause is always structurally superfluous; it plays no role whatever with respect to the main verb and no morphemes in the language are marked as requiring it. A relative clause sentence is equivalent to two independent predications on the same argument. These differences are captured by an analysis in which sentential subjects and objects are instances of underlying embedding, and relative clauses are only superficially embedded. If relative clause sentences are not underlyingly embedded structures this could account in part for the general disagreement, pointed out in (b) above, as to the underlying representation of the position of the embedded sentence.

II. The derivation of relative clause sentences.

A. Assumptions

In order to present the schematic outline for forming relative clause sentences, two assumptions must be made explicit.

(a) The difference between parts of sentences such as the following:

(3) I know a student who plays the harmonica.
(4) I know the student who plays the harmonica.

will be assumed to be introduced at some level of derivation other than the one at which "content morphemes" and the relations among them are specified. I leave open the question of just where such a distinction must be made; for the present discussion, it suffices to point out that (3) and (4) must have identical representations insofar as the meanings of the nouns and verbs and the relations among them are concerned. I shall further assume that the choice of the definite determiner will in general correlate with certain presuppositions which the speaker makes about the extent of his listener's knowledge.

(b) As pointed out by Bach (1968), numerals and quantifiers must be introduced outside the clause in which they ultimately appear. That this must be so is illustrated by the fact that the sentences of (5) are not matched by the respective pairs in (6):

(5) a) I have three students who are flunking.
b) I know few people who smoke cigars.
c) I saw no students who had short hair.

(6) a) { I have three students.
Three students are flunking.
b) { I know few people.
Few people smoke cigars.
c) { I saw no students.
No students had short hair.

B. Derivation

Returning now to the proposal for deriving relative clause sentences from conjunctions, I suggest that underlying (7) is a structure like (8):

(7) I met the girl who speaks Basque.
(8) (I met girl) (girl speaks Basque)

The choice of the clause to become the relative clause correlates with certain presuppositions on the part of the speaker about what the hearer knows, and accordingly with the choice of the determiner. Consider (8) again. If the speaker presupposes that the hearer knows neither about his meeting a girl nor about a girl's speaking Basque, then both of the following conjunction realizations of (8) are acceptable:

(9) I met a girl and she speaks Basque.
(10) There's a girl who speaks Basque and I met her.

as well as both of the following relative clause sentences with indefinite head nouns:

(11) I met a girl who speaks Basque.
(12) A girl I met speaks Basque.

If, on the other hand, the speaker presupposes that there is a girl such that it is known by the hearer that he met her, the relative clause sentence corresponding to this presupposition will have the conjunct containing <u>met</u> as the relative clause, and the head noun will be definite:

(13) The girl I met speaks Basque.

Similarly, if the speaker presupposes that his hearer knows about the girl who speaks Basque, the corresponding relative clause sentence will have the conjunct <u>speaks Basque</u> as the relative clause, and again the head noun will be definite:

(14) I met the girl who speaks Basque.

C. Implications

(a) The distinction then, between the "matrix" and "constituent" sentences in a relative clause structure can be seen to be related to nothing in the structural portion of the representation of such sentences. The meaning difference between sentences (13) and (14), in other words, is not a function of the fact that the matrix and the constituent sentences have been interchanged; if it were, then we should expect the same meaning difference to characterize the pair (11) - (12). But (11) and (12) do not have different meanings in any usual sense of the word "meaning". Instead, the semantic difference between (13) and (14) is a function of the presuppositions which the speaker has about the extent of his hearer's knowledge.

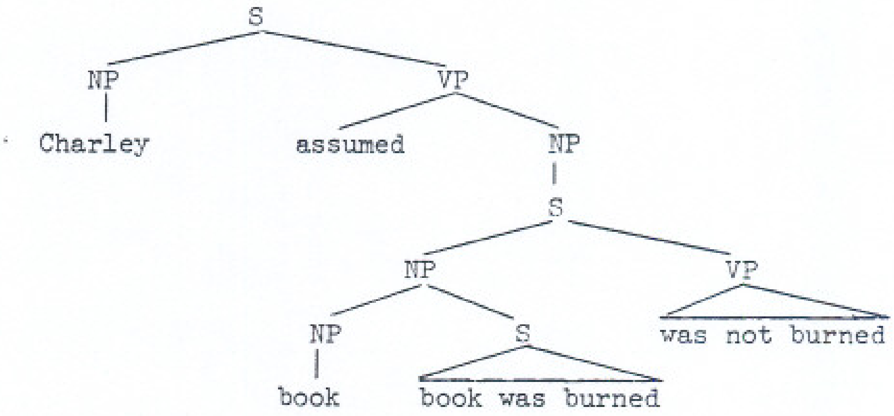
(b) Similarly, the "restrictiveness" of a relative clause is also shown not to be a property best described in terms of an embedding underlying representation. Relative clauses with indefinite nouns do not "restrict" these nouns in the way that relative clauses with definite nouns seem to, and yet underlying embedding structures do not reveal a basis for this difference. Again, I think that the apparent "restricting" nature of relative clauses with definite head nouns is a function of the presuppositions discussed above.

(c) Postal (1967) has shown that a certain ambiguity can be explained only if relative clauses are assumed to be derived from conjunctions. The sentence he gives is:

(15) Charley assumed that the book which was burned
was not burned.

On one reading, Charley assumed that a certain book had not been burned when in fact it had been. On the other reading, Charley assumed a contradiction. On the hypotheses that relative clause sentences are underlyingly embedding structures, there is no way to represent the ambiguity. This is because corresponding to (15), only one embedding structure can be constructed, namely:

(16)



But there are two conjunction sources for (15). Underlying the first reading, in which Charley is merely mistaken, is the representation:

(17) ((Charley assumed (book not burned)) (book burned)

Notice that, as we would expect, (17) also underlies:

(18) The book which Charley assumed was not burned was burned.

which results from the first conjunct's becoming the relative clause, as well as the conjunction:

(19) Charley assumed that the book was not burned but it
was burned.

Underlying the second reading, in which Charley assumes a contradiction, is:

(20) Charley assumed ((book burned) (book not burned))

As with (19), (20) underlies two sentences besides (16). By selecting the second of the two conjuncts of (20) as the relative clause, we can derive:

(21) Charley assumed that the book which was not burned
was burned.

which is an exact paraphrase of the second reading of (15). The conjunction derivable from (20) is, of course:

(22) Charley assumed that the book was burned and that
it was not burned.

At this point, it should be made clear that there is one class of relative clause sentences which do not seem to be related to conjunctions in the manner just described. A sentence such as:

(23) Men who smoke pipes look distinguished.

which contains a relative clause with a generic head noun, obviously does not have a conjunction such as:

(24) (men smoke pipes) (men look distinguished)

as its source. It is generally assumed that such a sentence is instead derived from the representation underlying an _if-then_ sentence like:

(25) If a man smokes a pipe, he will look distinguished.

The extremely interesting semantic and syntactic issues raised by this assumption will unfortunately be left unexplored here.

III. Non-restrictive relative clauses.

The similarities between non-restrictive clause (=NR) sentences and conjunctions have been remarked upon by a number of linguists (see, for example, Annear (1967), Drubig (1968), Lakoff (1966), Postal (1967), Ross (1967)). I will not review these similarities, but I will assume that NR sentences must be derived from conjunctions. Again, as far as I know, no arguments have been advanced in favor of an embedded analysis for NR sentences; in those studies which present underlying embedding representations for NR's, the question of there being alternative analyses is not even raised.

At the outset, two types of NR sentences must be distinguished; I will refer to them as Type I and Type II NR sentences. Type I NR sentences are exemplified by:

(26) Jerry, who used to play football, now has a
sedentary job.
(27) I had a date with the librarian, who read to me
all evening.

Type II NR sentences are exemplified by:

(28) She took the children to the zoo, which was very
helpful.

(29) Joe debated in high school, which Chuck did too.

In type I NR sentences, the relative pronoun replaces a referring noun phrase; in Type II, it replaces an entity, the nature of which will be clarified later in this section. For the moment, we will consider only Type I.

A. Type I NR's

Ross' proposal (1967, p. 174) that all Type I NR's be derived from second conjuncts seems to be correct. That is, at some intermediate level before anaphoric pronominalization has applied, given a conjunction each of whose clauses contain an occurrence of a co-referential noun, the second conjunct can be moved to a position immediately following the noun in the first conjunct. Pronominalization can then apply, moving either backwards or forwards[1], so that from the conjunction

(30) George noticed that Margie refused the candy, and George didn't take any candy.

any of the following can be derived:

(31) George, who didn't take any either, noticed that Margie refused the candy.
(32) George, who noticed that Margie refused the candy, didn't take any either.
(33) George, who didn't take any candy, noticed that Margie refused it too.
(34) George, who noticed that Margie refused it too, didn't take any candy.

One apparent counterexample to the claim that NR's are derived from second conjuncts is the following sentence:

(35) Is even Clarence, who is wearing mauve socks, a swinger?

As Ross (1967) points out, its conjunction counterpart does not exist:

(36) *Is even Clarence a swinger, and he is wearing mauve socks?

It seems to me that Ross' solution to this problem is not as radical as he indicates. As a source for (36) he proposes the structure underlying:

(37) Is even Clarence a swinger? Clarence is wearing mauve socks.

[1] Ronald Langacker pointed out this fact to me.

Instead of following Ross in his conclusion that all NR's must be derived from sequences of sentences, I claim instead that the connector is deleted between a question and a declarative.

Imperatives are similar to questions in this respect. The source of:

(38) Tell your father, who is outside, that supper is ready.

apparently cannot be:

(39) *Tell your father that supper is ready, and he is outside.

But if there is a rule deleting _and_ between imperatives and declaratives, the problems disappear. Notice that it would not help to posit a conjunction source in which the declarative sentence came before the question or imperative; questions and imperatives simply cannot be connected to declaratives by _and_, either before them or after them.

(40) Clarence is wearing mauve socks, and is even he a swinger?

(41) Your father is outside, and tell him that supper is ready.

Finally, a restriction must be placed on the NR rule to the effect that questions and imperatives themselves cannot become NR's.

At this point two objections might be raised; I would like to consider these in slightly greater detail. First, it has often been suggested that an NR represents an assertion by the speaker, a comment injected into the sentence whose truth is being vouched for by the speaker independently of the content of the rest of the sentence. An example of the type of sentence which makes such an analysis seem likely is

(42) The mayor, who is an old windbag, designated himself to give the speech.

An implication of this analysis is that NR sentences should be represented in such a way as to reflect that the NR is an independent assertion made by the speaker, perhaps by positing a separate superordinate declarative performative for it. However, it is not correct to assign the responsibility for the truth of every NR to the speaker of the sentence in which it occurs. Bach (1968, p. 95) points out that a sentence like

(43) I dreamt that Rebecca, who is a friend of mine from college, was on the phone.

which might be thought to contain an NR asserted by the speaker, can be made ambiguous by changing _is_ to _was_. The case is even clearer in a sentence in which the subject is different from the speaker.

It seems to me that the following sentences are ambiguous as to whether the subject or the speaker is vouching for the truth of the NR:

(44) Harold says that his girlfriend, who is a little bit crazy, wants to go to Hanoi.
(45) The claims agent said that the paint job, which should have been done long ago, would cost $150.

In fact, each of the above sentences can be disambiguated by adding a clause which forces the interpretation in which it is the subject, rather than the speaker, who asserts the NR.

(46) Harold says that his girlfriend, who is a little bit crazy, wants to go to Hanoi, but I think she's too rational to try it.
(47) The claims agent said that the paint job, which should have been done long ago, would cost $150, but he doesn't know that now is when it should be done.

The other possible objection to my thesis is that if both non-restrictive and restrictive relative clause sentences are derived from conjunctions, then sentences of both types, which may have very different meanings, can be derived from identical sources. Arguments against having identical sources for the two types of sentences carry weight only for sentences with numerals in them, which I will discuss shortly. In other cases, it seems that once again the differences between restrictive and non-restrictive relative clause sentences are not of the sort that ought to be represented structurally; instead they are differences representing a speaker's decision about how to present to the hearer information present in the underlying representation. For example, consider the two sentences:

(48) The boy, who works at the library, is majoring in philosophy.
(49) The boy who works at the library is majoring in philosophy.

The representation underlying both of these is:

(50) (boy works in library) (boy is majoring in philosophy)

For (48) the speaker has decided that the boy is already known to the hearer; the speaker is adding two pieces of information about that boy. For (49) the speaker assumes that the hearer knows about the boy who works at the library; <u>the</u> can be used with this NP, and the information which the speaker assumes to be new appears as the main predicate. I can see no way in which such a difference as that which exists between restrictives and non-restrictives could be represented in a consistent way for all such sentences in terms of some underlying structural distinction.

Restrictive and non-restrictive relative clause sentences with numeral associated with the head nouns do have different representations. Consider the sentences:

(51) Three boys who had beards were at the party.
(52) Three boys, who had beards, were at the party.

The assertions are quite different: (51) means not that three boys were at the party, but that there were three boys all of whom both attended the party and had beards. But (52) does mean that there were three boys at the party. Understanding very little about the representation of numerals, I can do no more now than to suggest that underlying (51), the numeral is associated with neither of the conjuncts, while underlying (52) it appears in both. This is confirmed by the fact that corresponding to (51) there is no two-clause conjunction, but corresponding to (52) we find:

(53) Three boys were at the party, and they had beards.

B. Type II NR's

Type II NR's are also derived from second conjuncts. The examples given above of Type II NR's were

(28) She took the children to the zoo, which was very helpful.
(29) Joe debated in high school, which Chuck did too.

I suggest that these are immediately derived from the sentences

(54) She took the children to the zoo, and that was very helpful.
(55) Joe debated in high school, and that Chuck did too.

Before outlining the process by which Type II NR's are formed, let us consider a derivation in reverse, with (28) as an example. Its immediate source is (54). The that of (54) is a pro-form for certain repeated portions of a sentence; directly underlying (54) would be

(56) She took the children to the zoo, and her taking the children to the zoo was very helpful.

Disregarding the tense of the first conjunct, we can see that the that in (54) has replaced the repeated portion of the second conjunct of (56). Let us take a derivation in reverse with another example:

(57) They said she could play the marimba, which she can.

The sentence containing that which immediately underlies (57) is

(58) They said she could play the marimba, and that she can.

Directly underlying (58) is the full form with the repeated portion preposed:

(59) They said she could play the marimba, and play the marimba she can.

The immediate source for (59) is

(60) They said she could play the marimba, and she can play the marimba.

In detail, the derivation of a type II NR sentence proceeds as follows: Given a near-surface-level conjunction in which part of the surface VP of the first conjunct matches part of the VP of the second conjunct, (a) the repeated portion may be preposed;[2] (b) the preposed portion may be replaced by *that*;[3] and (c) the connector may drop, with concomitant change of *that* to *which*.

Notice that, as outlined by Chomsky (1957), when there is no auxiliary element to carry emphasis or negation, a *do* must be added, as in the following examples:

(61) She promised to dance for us, and she did dance for us.
(a) She promised to dance for us, and dance for us she did.
(b) She promised to dance for us, and that she did.
(c) She promised to dance for us, which she did.
(61) She dances well, and I don't dance well.
(a) She dances well, and dance well I don't.
(b) She dances well, and that I don't.
(c) She dances well, which I don't.

The following examples show the operation of an optional rule of "parenthesis:"

(63) That Cornelius was pleased was to be expected, and he certainly seemed to be pleased.
(a) That Cornelius was pleased, and he certainly seemed to be pleased, was to be expected.
(b) That Cornelius was pleased, and pleased he certainly seemed to be, was to be expected.
(c) That Cornelius was pleased, and that he certainly seemed to be, was to be expected.

[2] This formulation is slightly inaccurate. Exactly what gets preposed will be described more carefully below.

[3] The order of these two rules will be reviewed below.

(d) That Cornelius was pleased, which he certainly seemed to be, was to be expected.

A special set of examples is the following, in which a *do* appears:

(64) She taught me to bake a cake, and I couldn't bake a cake before.
(a) She taught me to bake a cake, and bake a cake I couldn't do before.
(b) She taught me to bake a cake, and that I couldn't do before.
(c) She taught me to bake a cake, which I couldn't do before.
(65) We read *Tom Sawyer*, and we had never read *Tom Sawyer* as children.
(a) We read *Tom Sawyer*, and read *Tom Sawyer* we had never done as children.
(b) We read *Tom Sawyer*, and that we had never done as children.
(c) We read *Tom Sawyer*, which we had never done as children.

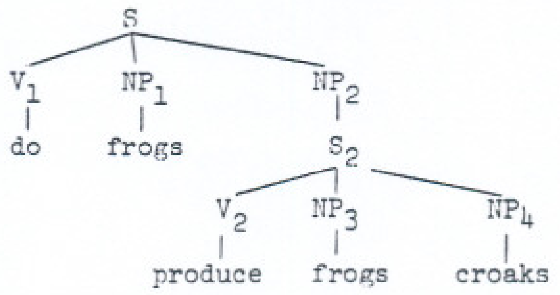
Sentences such as (64) and (65), when considered with certain other sentence types, provide evidence for two related hypotheses.

The first, advanced by Ross, is that activity verbs are associated at some level with the "primordial" action verb, *do*.[4] I understand him to be claiming that this *do* is present in the underlying representation of all activity sentences. Because its occurrence is entirely predictable, I would choose not to view it as present at this level, but as inserted into activity sentences early in their derivation.

The second hypothesis which sentences such as (64) and (65) provide evidence for is that the *do* in such sentences has as its object an NP. According to Ross, the NP in question is the underlying object of *do*, and it is an entire sentence:

(66) Frogs produce croaks.

[4] I cannot fully appreciate Ross' position since I have access to it only in the very sketchy form of a handout from his paper, "Act," presented at the July 1969 meeting of the Linguistic Society of America in Urbana, Illinois. From this handout, and from reports on the paper, I believe that the points which I have attributed to Ross are accurately stated here.



Aside from the fact that there seems to be no evidence for $NP_3$, that is a second underlying occurrence of the surface subject, the evidence which indicates that the *do* must take an NP object indicates that it is not an underlying NP that we are concerned with here at all, and that it is not a sentence. Let us consider this evidence. In a sentence like

(67) I realized that Art had visited the Dean, which I should do too.

we are tempted to declare that the *which* replaced an NP, since we know that in restrictive relative clause sentences and in Type I NR sentences, *which* always replaces an NP. However, this is not a very strong argument, since in questions, *which* can replace a demonstrative:

(68) Which book did you steal? I stole this book.

But the argument that *which* replaces an NP becomes more convincing when we consider the immediate source for (67), namely:

(69) I realized that Art had visited the Dean, and that I should too.

Beyond these NR sentences, no example of *that* replacing anything but an NP comes to mind. Further support comes from a paraphrase of (67):

(70) I realized that Art had visited the Dean, (which is) something I should do too.

*Something* is the NP pro-form *par excellence*, and it is clearly the object of *do*. But what it is coreferential with is not the sentence:

(71) Art had visited the Dean.

since what underlies sentences (67), (69), and (70) is not

(72) *I realized that Art had visited the Dean, and I should Art visit the Dean too.

What underlies (67), (69), and (70) instead is

(73) I realized that Art had visited the Dean, and I
should visit the Dean too.

In other words, somehow the phrase visit the Dean must be an NP before the rules changing this phrase to that apply.

Ross has suggested that pseudo-cleft sentences provide additional support for the hypothesis that phrases like visit the Dean must be NP's:

(74) What I should do is visit the Dean.
(75) Art did what I should do: visit the Dean.

What examples (67) through (75) show is that the NP which the NR and pseudo-cleft rules, and certain other rules, must refer to need not be an S at any level.

Further evidence that the NP referred to by these rules is a surface NP rather than an underlying NP can be found in the fact that what follows surface be must also be an NP. A collection of relevant examples is

(76) Nick is tall, which I will never be.
(77) Nick is tall, (which is) something I will never be.
(78) What I will never be is tall.
(79) Nick is what I will never be: tall.

Ross (1969) has used examples like these to show that adjectives must be underlying NP's. However, examples like the following show that adjectives and other post-be expressions must be not underlying but superficial NP's.

(80) I saw that Irma was easy to please, which I should
be too.
(81) I saw that Irma was easy to please, (which is)
something I should be too.
(82) What I should be is easy to please.
(83) Irma is what I should be: easy to please.

The expression easy to please in (80) - (83) cannot be an underlying NP, since in deep structure easy and please are not even constituents of the same S:

(84) ((one please Irma) easy)

In the examples

(85) Chinese was easily mastered by Rich, which it was
not by Claire.
(86) Chinese was easily mastered by Rich, (which is)
something it was not by Claire.

(87) What Chinese was was easily mastered by Rich.

We can see that the phrase *easily mastered* is not an underlying complement of *be* for there is no underlying *be*; moreover, since the verb *master* is an activity verb, at some intermediate level it would actually be the object of *do*.

My proposal, then, is the following: neither *do* nor *be* is present in underlying representations. *Be* may become the main verb by any of a variety of well-known obligatory transformations. *Do* is inserted preceding activity verbs. At the point at which *do* or *be* is inserted into a sentence, the part of the VP which follows becomes an NP; its NP status is then referred to by a number of optional rules, such as those which produce the sentences we have been considering here. If none of these rules applies to separate the *do* from its object, Ross' rule of '*do*-gobbling' applies, deleting *do's* that are directly followed by their objects.

If this analysis is in general correct, we are ready to reformulate the steps by which Type II NR's may be formed. Rephrasing the set of three rules (a) - (c) given earlier, we arrive at the following statement: Given a near-surface-level conjunction in which part of the surface VP of the second conjunct is a repetition of part of the surface VP of the first conjunct, (a) the NP "complement" of *be* or *do* may be preposed; (b) this NP may be replaced by *that*; and (c) the connector may drop, with concomitant change of *that* to *which*. This reformulation corrects two inaccuracies in the previous (a) - (c). The earlier formulation said that the portion of the second conjunct involved in these rules was the "repeated portion." This is not quite accurate, since in

(83) Nick is tall, and I shall never be tall.

*be* is part of the repeated portion of the second conjunct (with tense disregarded). But clearly the *be* is not part of what is changed to *that*, or preposed:

(89) Nick is tall, which I shall never be.
(90) Nick is tall, and that I shall never be.
(91) Nick is tall, and I shall never be that.
(92) *Nick is tall, which I shall never.
(93) *Nick is tall, and that I shall never.
(94) *Nick is tall, and I shall never that.

What does achieve the desired results is the requirement that what is preposed or changed to *that* be an NP.

Second, the order of rules (a) and (b) is irrelevant now, since *that* can appear either after *do* or *be* or in its preposed position. Beginning with the initial sentence of (64), we derive

(95) She taught me to bake a cake, and bake a cake I
couldn't do before.

by applying (a) alone,

(96) She taught me to bake a cake, and I couldn't do that before.

by applying (b) alone, and:

(97) She taught me to bake a cake, and that I couldn't do before.

by applying both rules. Similarly, beginning with (80), we derive

(98) I saw that Irma was easy to please, and easy to please I should be too.

by applying (a) alone, and:

(99) I saw that Irma was easy to please, and I should be that too.

by applying (b) alone, and:

(100) I saw that Irma was easy to please, and that I should be too.

by applying both rules.

One final minor point. A _do_ occurring immediately after a stressed modal may be dropped. Thus, sentences (57) and (65) have a variant form with final _do_:

(101) They said she could play the marimba, which she can (do).

In this section I have considered two types of NR sentences, showing how both are related to near-surface conjunctions, and how NR sentences of Type II provide evidence for two hypotheses, one that activity sentences have at some level _do_ as main verb, and the other that only at a fairly superficial level must the phrase following _do_ or _be_ be an NP.

IV. Summary.

I have tried to present some heretofore unexamined evidence that both restrictive and non-restrictive relative clauses must be derived from underlying conjunctions, and that this can be achieved in a grammar with certain well-motivated and fairly traditional restrictions on what aspects of the meaning of a sentence are to be represented at the structural level of its underlying representation.

## APPENDIX

As this paper was going to press, a squib appeared in Linguistic Inquiry 1.3, July 1970, by David Perlmutter and John Robert Ross, in which it was proposed that sentences like

(i) a man entered the room and a woman went out who
were quite similar

"present the theory with a new paradox." In their words,

> Neither of these singular noun phrases can serve as the antecedent of a relative clause whose predicate (similar) requires an underlying plural subject, and whose verb (were) is inflected to agree with a plural subject in surface structure. The only possible antecedent of the relative clause in (i) would seem to be the discontinuous noun phrase a man ... (and) a woman. But how can a discontinuous noun phrase be the antecedent of a relative clause? No analysis of relative clauses that has yet been proposed for the theory of generative grammar is able to account for sentences like (i). (p. 350).

I would like to suggest that sentences such as (i), [which are indeed anomalous in a traditional embedding analysis of relative clause sentences], present no paradox at all if relative clause sentences are viewed as underlying conjunctions; the conjunction source for (i) would simply be:

(ii) (man entered room) (woman went out of room)
(man and woman were similar).

References:

Annear, Sandra S. See Thompson.

Bach, Emmon (1968) "Nouns and noun-phrases," in Bach and Harms (1968).

Bach, Emmon and Robert T. Harms (1968) *Universals in Linguistic Theory*, New York: Holt, Rinehart and Winston.

Chomsky, Noam (1957) *Syntactic Structures*, The Hague: Mouton.

Chomsky, Noam (1965) *Aspects of the Theory of Syntax*, Cambridge, Mass.: M.I.T. Press.

Drubig, Bernhard (1968) "Some remarks on relative clauses in English," *Journal of English as a Second Language* 3.2.23-40.

Fillmore, Charles J. (1968) "The case for case," in Bach and Harms (1968).

Jacobs, Roderick, and Peter Rosenbaum (1967) *Grammar I*, *Grammar II*, Boston: Ginn & Co.

Lakoff, George (1966) "Deep and surface grammar," Unpublished paper.

Langendoen, D. Terence (1969) *The Study of Syntax*, New York: Holt, Rinehart and Winston.

Postal, Paul (1967) "Restrictive relative clauses and other matters," Unpublished paper.

Reibel, David and Sanford Schane (1969) *Modern Studies in English,* Englewood Cliffs, N.J.: Prentice-Hall.

Ross, John R. (1967) "Constraints on variables in syntax," Ph. D. Dissertation, M.I.T.

Ross, John R. (1969) "Adjectives as noun phrases," in Reibel and Schane (1969).

Thompson, Sandra Annear (1968) "Relative clauses and conjunctions," *Working Papers in Linguistics* No. 1, Ohio State University Research Foundation.

UCLA English Syntax Project Report (1969) Los Angeles.

# Speech Synthesis Project*

David Meltzer

*Sponsored in part by the National Science Foundation through Grant GN-534.1 from the Office of Science Information Service to the Computer and Information Science Research Center, The Ohio State University.

# Speech Synthesis Project

David Meltzer

The purpose of this project has been to attach a terminal analog speech synthesizer of the Glace-Holmes type (JAWORD) to the PDP-10 computer to allow on-line speech synthesis in an interactive environment.

The JAWORD synthesizer is an advanced terminal analog type with 7 resonators and two forms of excitation (see Figure 1). The frequency and amplitude of each resonator are controlled by an externally supplied voltage of from 0 to +3 volts. Of the 14 parameters thus needed, only 10 are used at any one time; which of these 14 are used is determined by an 11th parameter which also controls which form of excitation is applied to the resonators (noise or pulse excitation). The important characteristics of this synthesizer from the viewpoint of interfacing it to a computer are:

1. Eleven analog voltages must be supplied simultaneously to the synthesizer.
2. A new set of voltages must be available every 10 milliseconds, i.e., the control voltages change at a very slow rate.

Previous experience had shown that the normal configuration of JAWORD did not allow sufficient control of the time of occurrence of the fundamental excitation pulses. The synthesizer was therefore modified so that the pulse could be supplied from an external source, in this case the computer.

The configuration used to control the synthesizer is shown in Figure 2. The digital control information generated by the control program is fed to the synthesizer controller via the PDP-10 Input-Output Bus. This controller, designed and built at Ohio State University, converts this information into the required analog voltages and supplies them to the synthesizer. The controller (see Figure 3a) uses an analog multiplexor feeding a series of capacitive hold circuits to generate the 9 continuously variable outputs from the output of one Digital-to-Analog Converter (DAC). The control word (Figure 3b) includes the information to set the DAC as well as the address of the hold circuit which will receive the value. Although considerably slower than a one DAC pre-channel system, its response is adequate for this application. Additional bits are provided in the control word for control functions, including a bit for the fundamental excitation pulse (1=pulse, 0=no pulse).

The logic of the controller is implemented in Digital Equipment Corporation B-series discrete component logic, the same family as used in the PDP-10 central processing unit (CPU). This family was chosen for reasons of ease of interfacing to the CPU. The DAC is a high precision 10-bit unit also made by Digital Equipment Corporation. The analog hold and amplifier circuits as well as miscellaneous level conversion circuits were designed and built at Ohio State University. The hardware is completely operational on-line with the CPU.

The next major element in the system is the control program for the synthesizer. The program is the basic interface between the user and the synthesizer. This control program operates as a privileged user job within the time-sharing environment so that other jobs may continue to use the system during speech synthesis jobs. There is a definite random degradation in output speech quality due to the presence of other activity in the system during synthesis, but this is not a problem for trial synthesis runs. A future revision of the control program is planned which will optionally turn off the time-sharing for the duration of synthesis output and then return to time-sharing mode. The actual output is of short enough duration so that this will not cause appreciable response time degradation.

The control program as now in operation performs several functions necessary for effective interaction with the experimenter. The primary level controls the operation of the I/o bus and appropriately sequences the outputting of control words to the synthesizer controller. These control words are generated by the next level of program, the conversion routine which translates from a standardized code (Carlson, 1969) to the appropriate channel addresses and DAC level, adds offset and calibration data and packs the words into a table in CPU core memory. This table becomes the input to the first program.

The basic interface to the experimenter is one of the Teletype consoles attached to the CPU. It is anticipated that this function can be taken over by the CRT display at some future date but the lack of adequate systems programming support and hardware character generator feature on the display makes this a difficult task.

Synthesis of a speech sample using the current program involves the following steps.

1. Typing in the sample in coded form and in-core editing it to the experimenter's satisfaction. The program provides for selective display and alteration of portions of the parameter table.

2. Outputting the speech sample and then making needed corrections.

3. Dumping the table in coded form on paper-tape for later use. The paper tape so prepared may be used as input in place of step 1.

The speech output may also be recorded on a built-in tape recorder. Future program versions will allow dumping the output on DEC tape so that many words or phrases may be kept on one tape. At present, however, there are not enough DEC tape drums available to allow each user of the system to have one tape and still let the synthesis job have two, one for program storage and one for speech sample storage.

The program as described above with paper-tape storage is fully operational. Evolutionary changes are being made to improve the output quality as experience with the synthesizer grows. Synthesis parameter tables distributed through SOUGHS (Society of Users, Glace-Holmes Synthesizer) are being used to gain experience with the system.

## Reference

Carlson, W. A. 1969, "On the Establishment of a Standard Format for Exchange of Data for the Synthesizer Among the Users," (private communication, distributed through SOUGHS).

Figure 1. JAWORD SPEECH SYNTHESIZER (SIMPLIFIED)

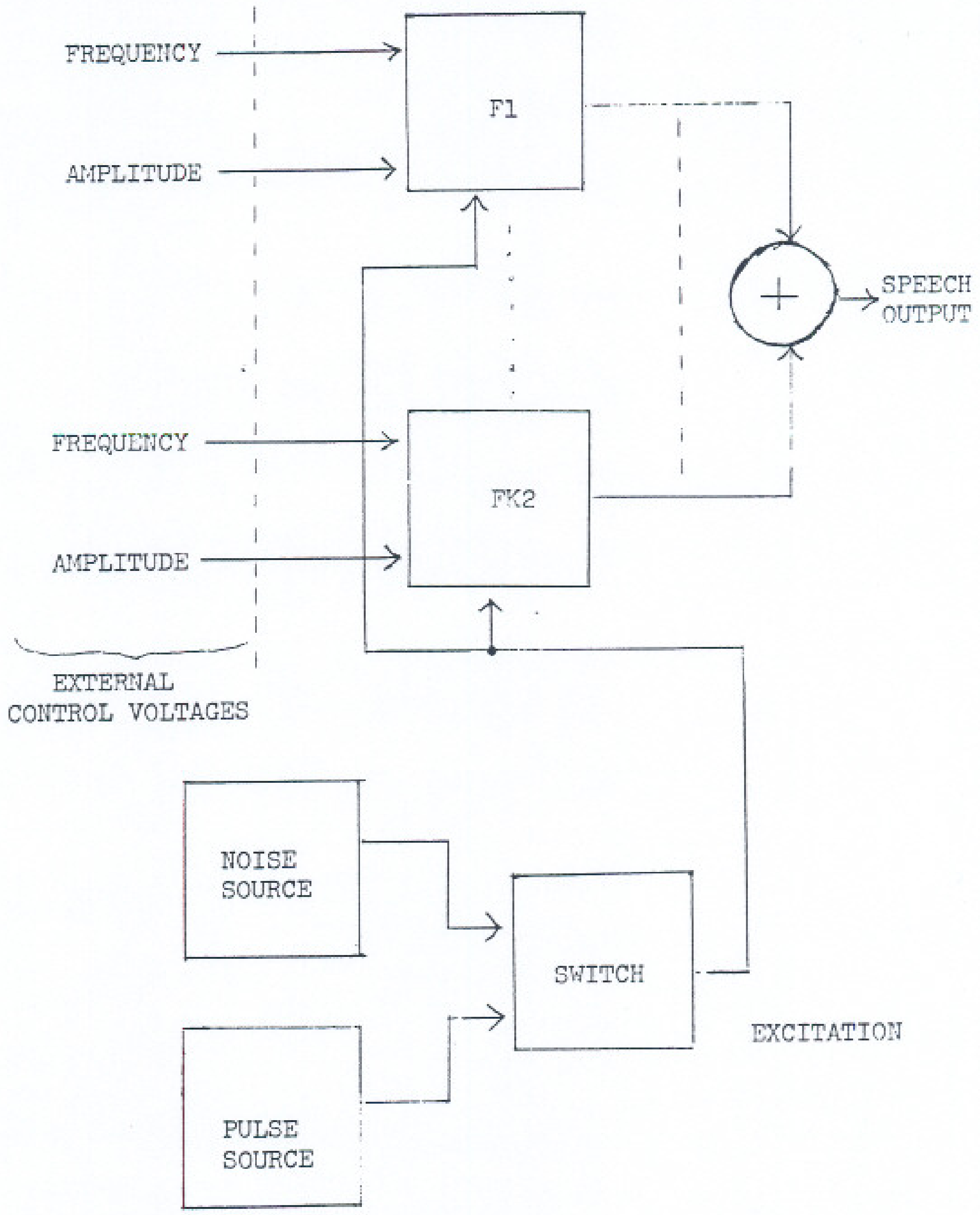
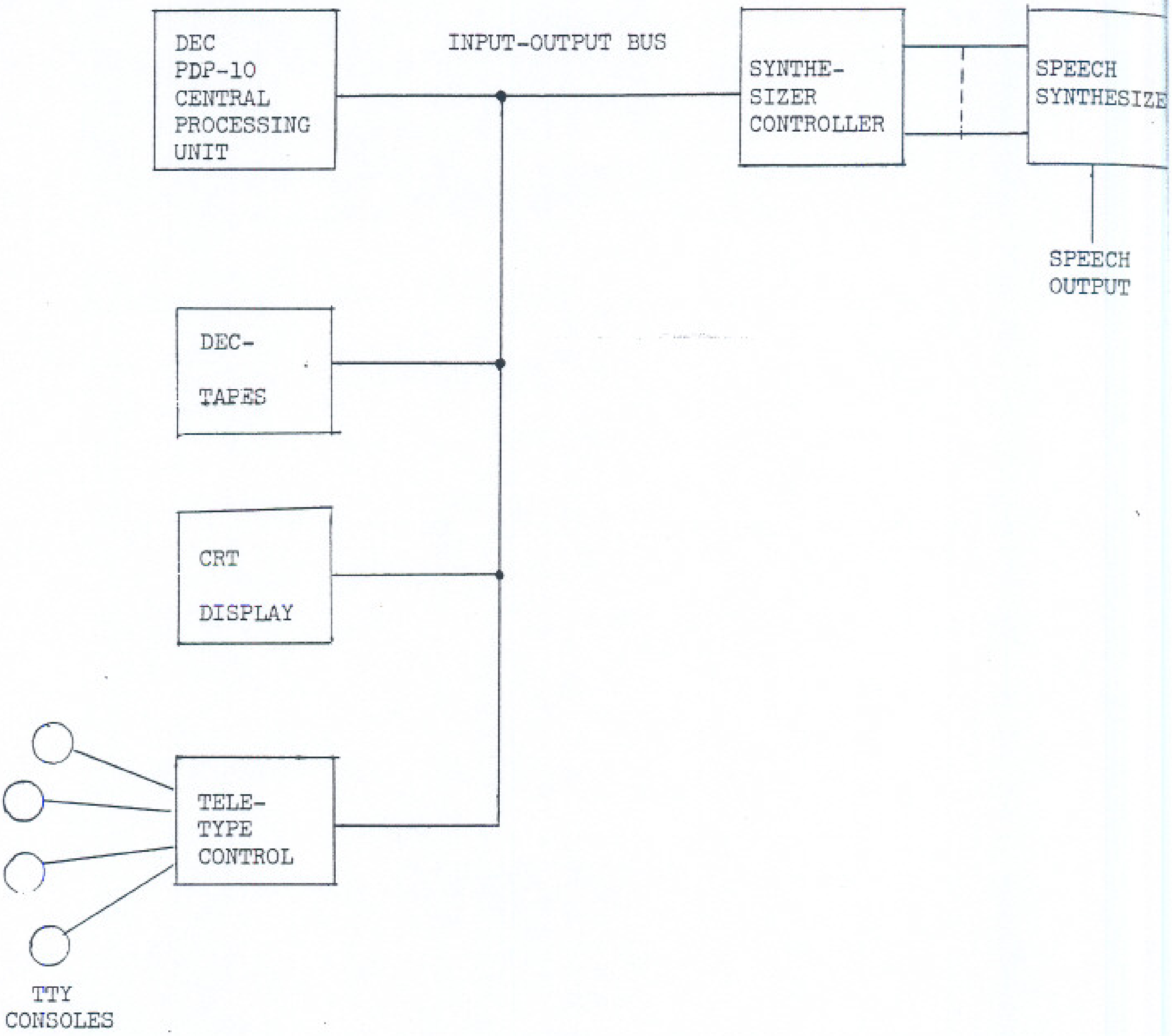
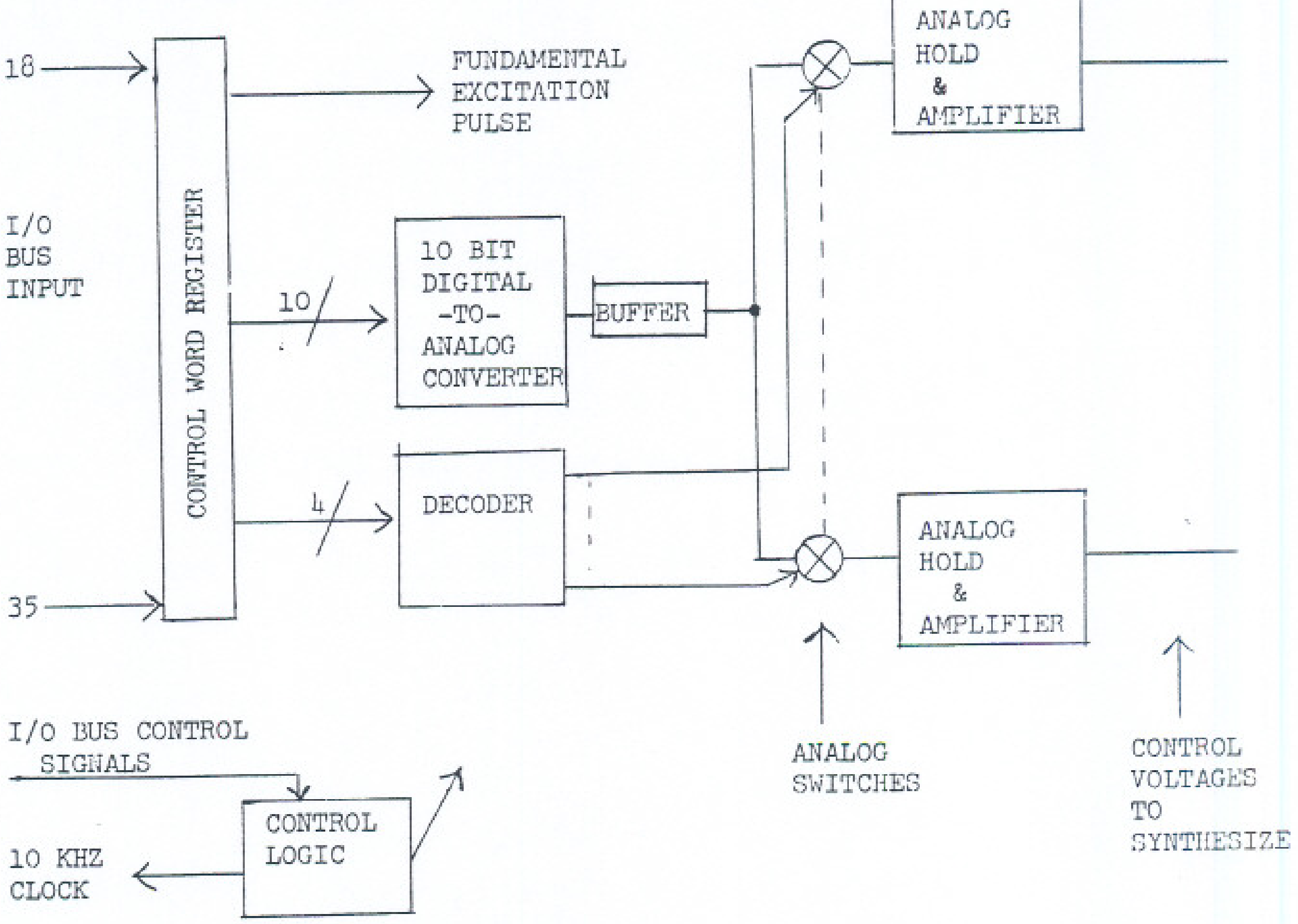


Figure 2. SYSTEM CONFIGURATION



## Figure 3. SYNTHESIZER CONTROLLER

### 3a. Functional Diagram



### 3b. Control Word Format

| 18 | 19 | 20 29 | 30 31 | 32 35 |
|---|---|---|---|---|
|  | F<br>Ø | DAC<br>VALUE | SWITCH<br>CONTROL | SCAN<br>ADDRESS |

← BIT POSITION IN MAIN MEMORY

# A Speech Production Model for Synthesis-by-rule*

Marcel A. A. Tatham

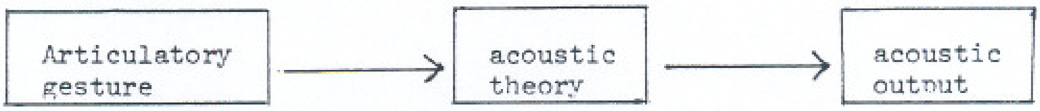
*Sponsored in part by the National Science Foundation through Grant GN-534.1 from the Office of Science Information Service to the Computer and Information Sciences Research Center, The Ohio State University.

# A Speech Production Model for Synthesis-by-rule

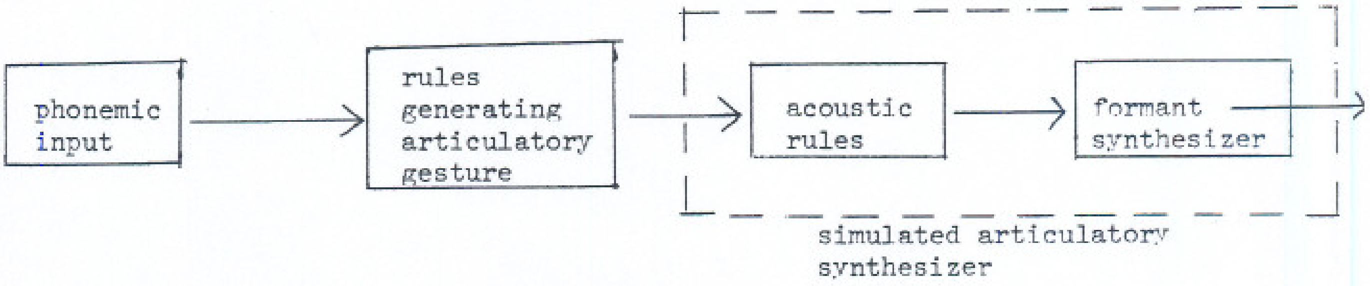
Marcel A. A. Tatham

## 1. Introduction

Problems in the design and operation of vocal-tract analog synthesizers suggest an interim handling of articulatory synthesis-by-rule. Relationships have been established between the acoustic output of speech and the articulatory gestures required to cause that output (Fant 1960, 1965; Flanagan 1965). These relationships are usually diagrammed:



As a preliminary then to articulatory synthesis the actual speech output from the device can be generated from a formant synthesizer which is immediately preceded by the rules of the acoustic theory, thus:


phonemic input
rules generating articulatory gesture
acoustic rules
formant synthesizer
simulated articulatory synthesizer

In the system described by Werner and Haggard (1969) a 'phonemic' input is turned into a number of time-varying articulatory parameters which correspond to spatial measurements in a stylized vocal-tract. A computer then applies a set of acoustic rules which are entirely extra-linguistic to generate from these articulatory parameters control signals which will operate a standard formant (or terminal analog) speech synthesizer.

Several very basic and influential assumptions underlie this approach, some of which reflect a non-linguistic viewpoint. Unlike a true vocal-tract analog, where in the ideal situation the sounds produced could not sound or be other than absolutely natural, this system relies on a terminal analog synthesizer. I have discussed elsewhere (Tatham 1970) some of the difficulties of synthesis in general and indeed it is noted in Werner and Haggard that the conversion tables are not free from 'tricks' (p. 5) which are designed to make the speech sound more natural. But a more striking criticism can be levelled at the use of terminal analog synthesizers.

Historically the parameters of terminal analog synthesizers and their control have been derived as a result of two criteria: (a) visual and (b) perceptual. (a) It is quite clear from the literature since 1950 that the visual inspection of spectrograms has dominated these choices (Lawrence 1953; Liberman et al. 1954). It was observed that in spectrograms vowel-like sounds exhibit two or three major formant areas in the frequency/amplitude domain: often a (possibly correct, but this is an 'after-the-fact' discovery) assumption was made: such obviously audible and therefore acoustically major parameters were clearly going to be the most perceptually relevant and ought therefore to be synthesized faithfully. (b) Later, relevance of individual parameters was established using perceptual experiments (Liberman et al., 1954). There is no doubt that perceptual criteria dominate approaches in terminal analog synthesis today. That a variety of stimuli can often produce similar perceptual responses cannot be denied and the absurd limit might be reached where for the sake of economy (of computer time, output interface, programming, bandwidth restriction in telephony, etc.) the output of the terminal analog will reduce even more its identity with real speech--yet sound similar.

These criteria of economy are never justified on linguistic grounds and particularly hard to defend in terms of speech production. That a 'nasal' can be perceived by juggling with formant amplitudes only clouds accurate modelling of the speech production system and may even bias perceptual experiments--the fact that the perceiver can be fooled is comparatively trivial.

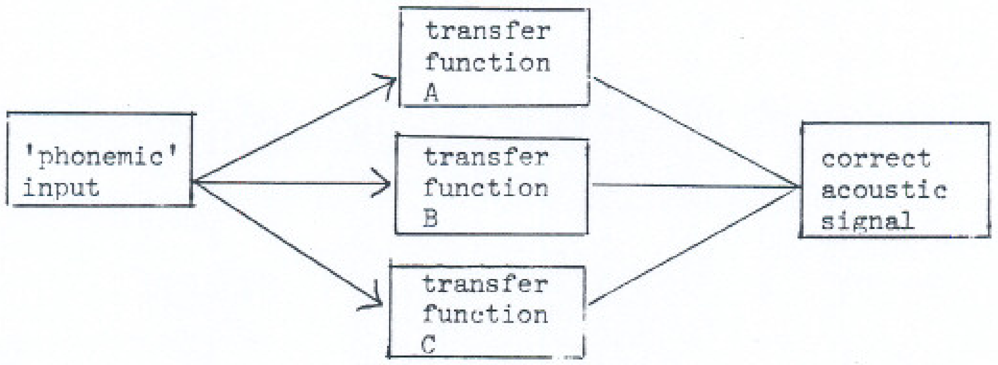
There is every reason to suppose that synthetic speech is a tool of considerable value in providing stimuli for perceptual experiments, but up till recently it has been a tool that may have been handled with more confidence than has been justified. It remains to be seen whether subjecting the listener to distorted artificial speech (rather than distorted real speech), as in the classical experiment by Broadbent and Ladefoged (1960), enables us to infer anything about the perception of real speech.

These of course are the reasons for preferring articulatory synthesis. As mentioned earlier vocal-tract analog synthesizers are still not entirely satisfactory--but there seems no reason to wait. The approach offered by the Cambridge group seems admirable: generate as an interim output articulatory parameters. Under ideal conditions these would be converted to parameter control signals for a vocal-tract analog, but pending this they can be converted to terminal analog control signals using the rules of the standard acoustic theory of speech production. If the speech is to be used for perceptual experiments then absolutely nothing is gained, however: properly the system should stop at the articulatory level.

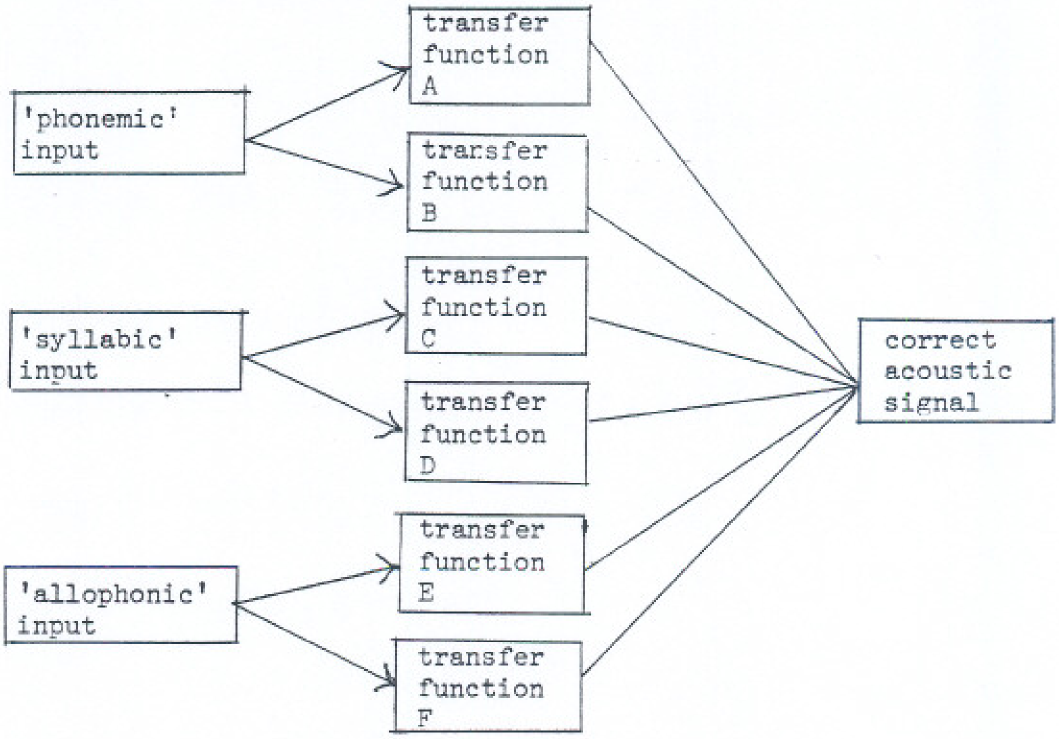
## 2. The Production Model

The present paper outlines a proposal for a model of speech production which is linguistically dominated and which might serve as the basis for an articulatory speech synthesis program. Linguistics has a lot to say about speech production which cannot be inferred from mere reproduction by lookup table of observed articulatory configuration, however. There is much more implied: it is not enough to generate correct vocal-tract shapes in an economical way for exactly the same reasons that it is not enough to generate 'correct' sounds. I have overstated my case deliberately; the incorporation as standard procedure these days of 'targets' and computed transitions based on contextual information is more than simply elegant--it does rest on linguistic theory: namely on the model which assumes that the phonological elements are phonemic (or quasi-phonemic) in nature and that allophones (or most of them) are the result of neuro-mechanical inertia at some low level in the system (Öhman, 1964). There is evidence that this view is inadequate however: the work of MacNeilage and Declerk (MacNeilage and Declerk 1968) has shown that there are segmental overlap phenomena not directly attributable to neural or muscular inertia (see also Tatham 1969a).

The synthesis strategy should be based on a coherent theory of speech production and the system described here will take into account one particular theory. But since there are competing theories a further function of the entire synthesis system will be to test the production theory. Notice, however, that the obtaining of a correct output is no test, just as the obtaining of correct perceptual response is no test of accurate acoustic signal. Consider, for example, the following possibilities:



In such a system there is no internal basis for evaluation of the model and the economy criterion mentioned above will not do as a measure of correctness. Consider the even worse situation:



Not only now is there no way of evaluating the various contending transfer functions but there is no internal way of evaluating the contending input types--but both these are candidates for evaluation and a measure of 'correctness' is crucial to the theory. Accordingly the transfer function must not simply be a mathematical formula which happens to provide the correct output from a particular input: there must be exterior constraints on this function.

Thus for example it is now established that muscle movement is not a continuously programmed system but a ballistic system controlled by temporally-spaced and non-continuous command situations. So at time T1 a muscle will be instructed to GOTO target t1 (this GOTO might be based on a given spatial movement, or on a given muscular tension), at time T2 an instruction will arrive to GOTO target t2 (a number of updating and correction signals may have arrived between T1 and T2)--but it is not the case that at T1 there is a 'start-moving' command and at T1+1 a 'move-a-little-more' command and at T1+2 a 'continue-moving' command and at Tt1 a 'right, -hold-it' command and at Tt1+1 a 'relax-a-little' command, etc.

Now what would these two different models of muscle command mean to the design of an articulatory speech synthesis system? The 'continuous command' theory would require the computation and supply of a moment-by-moment command signal at some level in the synthesis system corresponding to neural signals arriving at the muscles. The GOTO, ballistic theory would require a single computation (of the target value command) and the supply of just this single command (if necessary repeated for updating purposes) at the same level in the system.

The choice of the GOTO command in the synthesis strategy in the articulatory model presupposes a formula for the actual contraction of the muscle upon receipt of the full command--i.e., a factor expressing the inertia of the muscle in question. This factor is clearly derived at a level different from the level at which the instruction was computed (since it is a property of the muscle itself). But in addition it will also be argued that the computation could not have been accomplished without this piece of data.

For example, consider just one parameter of the muscle contraction: rate. In order to contract X amount, time Tx is necessary. Now, Tx will not alter the value (in a simple model) of command Cx, but it will alter its timing of delivery relative to the desired timing of achievement of contraction CNx. Thus a command signal Cx is computed based on at least two input channels: (a) the need for contraction of the particular muscle [command from a higher level] and (b) data about the rate of inertia of this muscle [data from a lower level].

Any synthesis strategy which began by observing in EMG or other data collection system that contraction for X began Tx before the achievement of X and simply arranged for this to be simulated would not even include the power of descriptive adequacy, let alone explanatory adequacy. But a strategy which includes the generalized data that this muscle is always (in the simple model) inert by a certain factor and computed a temporal change of the delivery of the instruction would provide explanatory adequacy--i.e., it would be able to predict in a transparent fashion the exact timing of the start of contraction of the muscle and would further provide the correct slope of the rate of achievement of that contraction.

In its simplest form then the model of speech production will have an input level to be equated with the input to the motor control system of human speech. It is not necessarily the case that this level is to be further equated with the level of systematic phonetics output from the phonological component of a transformational grammar--although this could be made to be the case.

Certainly there will be identity in the temporal respect. Phonology contains only a notional time: that of sequencing of segments. One immediate function of the production model is to transform this notional time into a less abstract time whose segments (whatever these should be decided to be) are organized on more than a sequential basis. Notice the important observation that such a model does not mention 'real-time'--indeed, it would be difficult to know what is meant by a 'real-time' model, in the correct sense of 'real'. Possibly a real-time model of speech production would deliver an output from an input in exactly the same time as a human being and with that time subdivided in exactly the same way as in the human being. Systems are said to operate in real-time, which means that there is no storage or slow-scanning of the data to make up for deficiencies in handling capacity in the simulator. Real-time notions do not affect the validity of the model or its ability to test the accuracy of its assumptions. It would be easy in advocating a performance model to misuse the term real-time; performance models merely have time other than notional time--they do not need to be real-time models or systems.

The segmental-input type will be assumed. There is enough psychological evidence for us to assume that there is a reality to segments. Strings will be assumed to be segmented in terms of extrinsic allophones--not in phonemes. This is nearly the case with all synthesis-by-rule systems. However our definition of segments will need to be different from that already established by researchers such as Mattingly (1968). It is not the case that the input segments will be phonemes except where the language has an idiosyncratic subdivision of phonemes which cannot be said to be co-articulatory (the classic example is: $L \rightarrow \left\{ \begin{array}{c} ɫ \\ l^{j} \end{array} \right\}$ in English). We will adopt the theoretical standpoint that rules of this type (properly allophonic rules that form part of the phonology, rather than the phonetics) have been applied to all segments. Thus, besides $L \rightarrow \left\{ \begin{array}{c} ɫ \\ l^{j} \end{array} \right\}$, we will also have X→x and Y→y, etc.; it is enough to argue this point on the grounds of symmetry alone, but we assume also that any phoneme is subject to a group of allophonizing rules, one function of which (<u>in the model</u>) is to switch levels of abstraction.

Thus the input to the model is characterized as a level expressed in terms of extrinsic allophones (Tatham 1969b). Recent experimental investigations of speech indicate an important and initial factor immediately influencing the ascription of time features to these segments. It seems to be the case that in C1VC2 utterances there is a motor-control link between C1 and <u>V</u> which cannot be explained by any low-level system or co-articulatory effect (MacNeilage and Declerk 1967; Tatham and Morton 1968b). <u>Where</u> this linkage or cohesion is introduced is not clear.

Notice the theoretical standpoint has been adopted that postulates that the cohesion has been introduced at a sub-phonological level. The point still needs to be argued in publication, but we will assume for the moment that although it is possible to construct a phonological component based on syllable segments (or segments of a similar kind) this is a theoretically clumsy and non-productive concept in abstract phonological theory. We shall assume (possibly wrongly--but decisions need be taken in a working model that complete the system; this is the difference between a working and a non-working model) that initial-CV cohesion is established at the motor-level. It is not crucial to the model (since it will satisfy the data without further speculation), but we might assume that the nature of the motor-control system is such that in speech this cohesion *must* be imposed.

Evidence from acoustic experiments (Lehiste 1970) supports the CV-cohesion theory, since these two elements remain non-compensatory--that is, complementary--under conditions of rate variation. The variation of rate is a factor to be accounted for crucially later. The data indicates that in cases of temporal strain on the overall word, compensation effects will occur between the V and C2 elements. This indicates temporal elasticity between V and C2 and temporal cohesion between C1 and V: that motor cohesion is observed (preceding paragraphs) is sufficient for us to introduce an actual linkage here which we could express with markers, thus:

| | |
|---|---|
| ≠C1V-C2≠ | where ≠ indicates a syllable boundary, juxtaposition (as C1V) motor-cohesion and-temporal compensation. |

operating as constraints in much the same way as +, ≠, etc., in the higher linguistic levels. The notation needs further thought because there will have to be rules deleting boundary symbols: these rules may have to be time-constrained. E.g., in cases of low rate speech we might well have

≠C1V-C2≠C3V-C4≠

where C2 and C3 are identical extrinsic allophones (e.g., 'black cat'); in high-rate speech we may want to add the rules:

| | | | |
|---|---|---|---|
| (i) | xC2≠C3y | xC5y | where C2 = C3 = C5 |
| (ii) | xCV-CVy | xCV≠CVy | where (i) and (ii) are ordered. |

Thus so far extrinsic allophones have been linked (for English) in two ways: motor-cohesion and temporal compensation. This composite linkage provides us with a complete syllable unit [≠C1V (-C2)≠] which still retains identity of its internal constituents. This is important because at this point there are two possibilities in the speech-synthesis strategy:

> look-up tables providing (initially) non-temporal (from the segment-sequencing viewpoint) information are consulted. These tables can be organized in one of two ways: (a) syllable-types are listed, (b) segment-types are listed.

(a) Each possible syllable type (we are concerned here only with the C1V part) is listed as a non-analyzable unit exhibiting two temporally-spaced GOTO targets. This will not be chosen because (i) (a theoretical reason) non-analyzability is rejected; (ii) data (often derived from slips-of-the-tongue experiments) indicate that the cohesion is not final. [note, however: slips-of-the-tongue experiments are confusing because there is often no evidence whether the slip has occurred at the phonological level (supports (a)) or at the phonetic level (supports (b)) (see, for example, Boomer and Laver (1967))].

(b) Segment types are listed together with an external set of rules (i.e. external to the segments) which determine cohesion. If cohesion is similar between all C1V possibilities this simply takes the form of a composite rule indicating in which motor parameters cohesion takes place and to what extent (NB the effect of this high-level cohesion on lower-level co-articulation, etc. will be discussed later). This solution satisfies the theoretical criterion of maximum generalization and compares favourably with the listing system of (a).

So far we have considered the characteristics of the input to the model and an initial stage intended to establish cohesions detected in experimental data. The theoretical model further assumes, as adjunct to the notion of GOTO control, that phenomena such as co-articulation are low-level rule-governed processes. These low-level processes are held to be true universals inasmuch as they reflect tendencies (predominantly inertial) of the neuro-muscular/mechanical system.

At this point it becomes necessary to discuss whether there is any attempt in higher-level programming to overcome such tendencies. So far, unfortunately, there is no definitive instrumental evidence, but it is assumed in the present model that (at least) a ternary system exists in the motor handling of (most of) the inertia-based effects. Inertia effects exist (they must, since all electrical or mechanical systems in the universe possess them)--the question is: are these effects handled in any systematic way; is any higher-level account taken of them? The ternary system in the model at the level postulates that one of three possible modifications exist: (i) counteract the effect, (ii) permit the effect, (iii) enhance the effect (these could be understood as -, 0, +, where 0 indicates the unmarked state). It is not clear from published data on co-articulation (Öhman, 1964, 1966, 1967; MacNeilage and Declerk 1968) (including over- and under-shoot) effects whether all mechanical or other inertia can be modified: presumably further data will be forthcoming; meanwhile the model will account, in the most simple way, for the existing data.

Thus, consider a language with only two palatal consonants of any one manner-type. Assuming a dominance of maximal differentiation (a psychological constraint) these will take the target forms of back and front (velar and alveolar, say), but this detail is comparatively unimportant. What is important is that the present model will predict a very wide variation in the point of contact of each consonant (but with little, if any, overlap) directly correlatable with segmental context. Thus preceding a front vowel, the consonant will exhibit a front allophone, etc. The model will further predict that this is the 0 or unmarked case--i.e., that there is no voluntary effort made to make the tongue less subject to context effect.

The model will predict, however, that, in another language where there are four such palatal consonants, (a) variation will again take place in exactly similar circumstances and (b) such variation will be very much more limited than in the case of the two-consonant language. The present model prefers to express this marked situation in precisely that two-level (or aspect) way *maintaining* the original inertia-derived rule and limiting it with a second, *linguistically-determined* rule. Thus the marking rule does *not* collapse two quite distinct and opposing tendencies--one quite a-linguistic and the other quite linguistic and concerned with maintaining perceptual clarity. Exactly the same phenomenon will be predicted for languages having a small number of distinctive vowel phonemes: the range of over-shoot and under-shoot variation will be considerable compared with a language with a larger number of vowels where the risk of perceptual confusion is that much greater if some kind of control is not exercised.

Notice that if control is to be exercised, the knowledge of the inertia effect must be possessed in *advance* by the control mechanism. This has got to be the case in this model; simple non-adjustable feedback systems cannot be relied on solely for one very simple reason [but there is an allowable alternative solution]: Language L1 with 3 palatal consonants and Language L2 with 5 palatal consonants both share a target value for one of their consonants--yet the range for L1 will be greater than the range for L2. But the model postulates a GOTO signal which will be identical in each case. Feedback cannot control the range of variation unless that feedback has been 'set' with respect to its limits: that such a possibility exists is well attested in the neurophysiological literature (see, for example, Matthews 1964). But the feedback cannot be set unless there is prior knowledge of the inertia that will occur and the steps that must be taken to contain the variation within the linguistically determined limits.

It could be argued that a relationship exists between the linguistics system and the bi-level inertia system such that it becomes language idiosyncratic to establish a relationship between the systems resulting in what has been termed an 'articulatory setting' (Drachman 1970). That there is a tonic state of the musculature (called 'basic-speech-posture') is undeniable and similarly that certain languages exhibit a predisposition for certain prevalent (usually secondary) phonetic characteristics (like velarization, retroflexion, predominance of lip-rounding, etc.). But we have only to discover one language with a small number of vowel phonemes with wide articulatory variation and at the same time with a large number of palatal consonants with a small degree of variation for this hypothesis to become suspect.

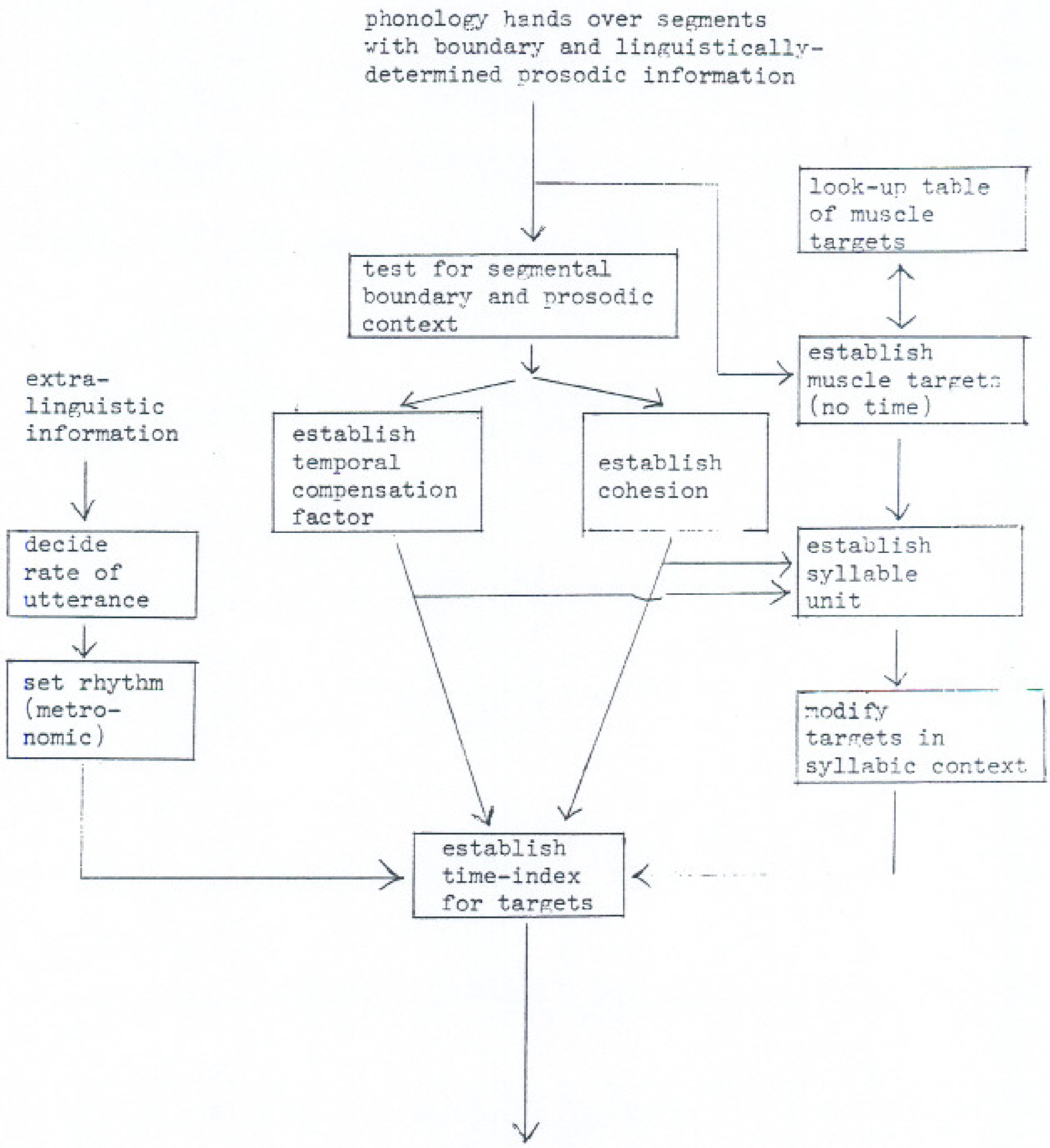
A second argument against this hypothesis is that it lends too much status to the low-level systems and gets them unsystematically involved in high-level phonological processes by postulating that phonological processes 'carry-along' with them arbitrary handling of the muscular and articulatory system.

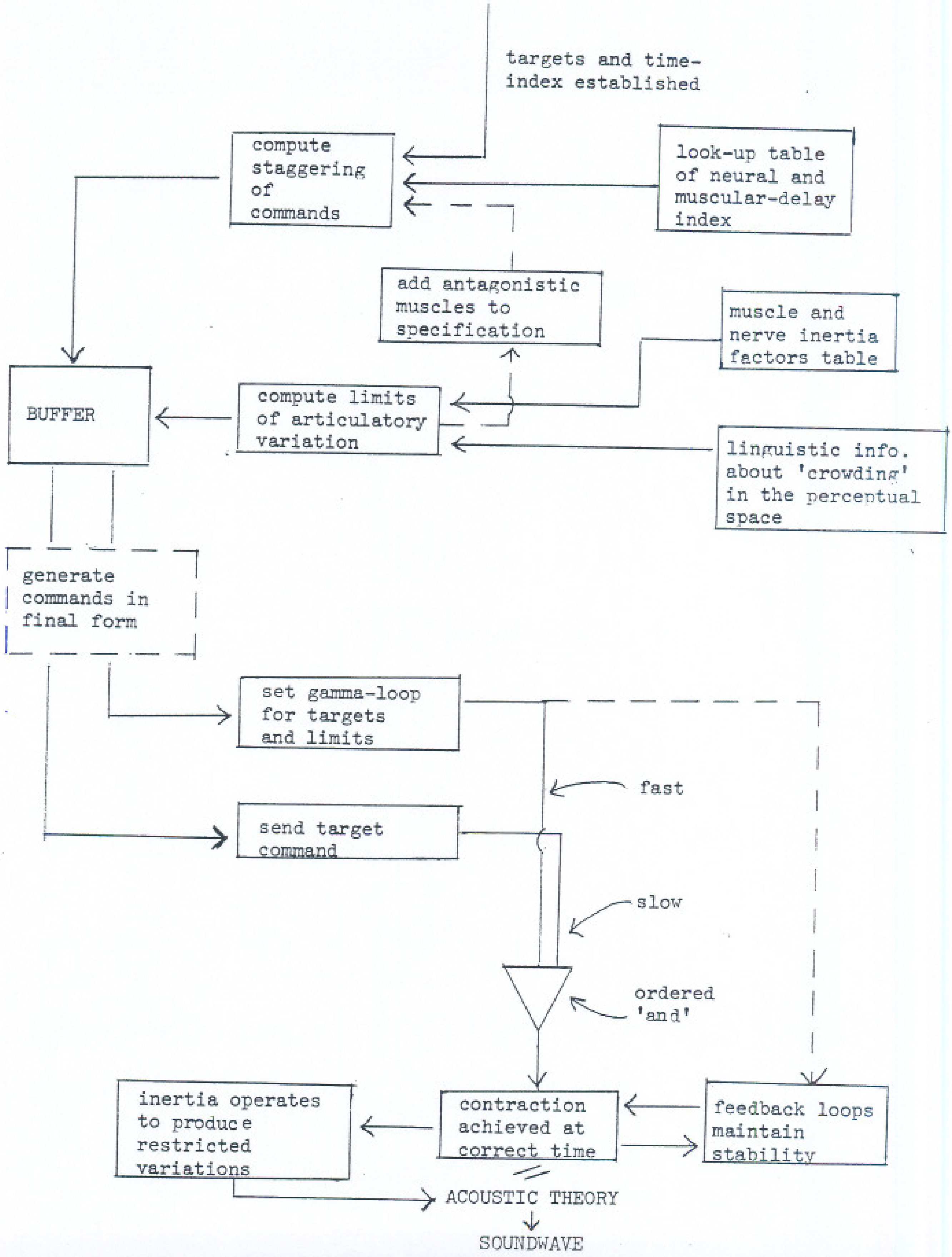
A third argument against this hypothesis is that it does not adequately account for the range of variation exhibited by a segment in a constant environment. If there were that much correspondence between mechanical inertia and phonology the articulation would be much more precise. The present model does predict a range of variation in the same segmental context because it only establishes *limits* for the variation, *not* new and different targets. I.e., the favoured model postulates a target, establishes the inertia formula, and establishes the limits to be imposed on that formula; the unfavoured model postulates a variety of a relationship established between mechanical tendencies and linguistic demands further resulting in an agreement for a particular and new target for each allophone [EMG records do not show that there is any contextual variation of this kind--but interpretation of such data is as yet only scantily formalized (Cooper 1965)].

Thus a particular articulatory gesture is the result of (a) the linguistically-motivated desire to articulate a particular extrinsic-allophonic segment, (b) the operation of a motor procedural mechanism establishing the cohesion of this segment with syllabic context, (c) the insertion of the composite syllabic-sized unit (of which this segment now constitutes a part) into the chosen rhythm or rate for this utterance, (c), in English, the modification of segmental duration depending on the stressed/unstressed pattern within the rhythm, (e) the generating of a target program associated upwards (i.e., linguistically) with this segment and horizontally (i.e., motor-wise) with the syllable unit, (f) the appendage of co-articulation limiting factors which limit the freedom of range of articulatory variables but which do not change the established target program.

Fig. 1 represents a simplified version of the present model. Some boxes are tentative (such as the setting of the gamma-loop system) but their *function* must occur somewhere to complete the system: they may just be in the wrong place or attributed to the wrong external mechanisms--what is correct about them is that, if included, then this model satisfies in a true explanatory way, the observables.

Fig. 1: TENTATIVE BLOCK DIAGRAM OF PROPOSED SPEECH PRODUCTION MODEL




targets and time-index established
compute staggering of commands
look-up table of neural and muscular-delay index
add antagonistic muscles to specification
muscle and nerve inertia factors table
BUFFER
compute limits of articulatory variation
linguistic info. about 'crowding' in the perceptual space
generate commands in final form
set gamma-loop for targets and limits
fast
send target command
slow
ordered 'and'
inertia operates to produce restricted variations
contraction achieved at correct time
feedback loops maintain stability
ACOUSTIC THEORY
SOUNDWAVE

Operations in Fig. 1

1. Input from the phonology decides which segment (=extrinsic allophone in sequential context with morpheme and word boundary symbols, stress pattern, etc.) is required at a particular point in the utterance to be generated.
2. Test for segmental context:
   (a) utterance (or inter-pause group): initial, medial, final?
   (b) (any) sub-utterance group: initial, medial, final?
   (c) word initial, medial, final?
   (d) morpheme initial, medial, final?
   (e) syllable initial, medial, final?
3. Establish whether motor-cohesion (lack of temporal compensation) or temporal compensation is to operate (this depends on the answer to 2).
4. (from 1) Look up muscle targets. If targets are expressed in terms of degrees of muscle contraction we have to ask the theoretical question: should targets for all muscles be specified irrespective of whether or not some are not involved in this particular segment, i.e., should a marked/unmarked system of classification be introduced? A combination of full entries with marked system would produce a segmentally determined feature hierarchy which is an issue of theoretical importance.
5. Establish relationship between targets and cohesion and compensation--i.e., establish syllable unit.
6. Decide overall rate of utterance.
7. Set rhythm generator according to 6: (i.e., provide metronomic determination).
8. Establish how each segment (incorporating 5) will behave temporally in the rhythm established by 7.
9. Hand over the information about the segment to the motor-command generator (this must be buffered to allow command-initiation overlap).
10. Establish information about any neural line-delay in the system.
11. Establish the muscle-response delay.
12. Construct a muscle command ordering dependent on 10 and 11 (at this point commands for individual muscles involved in the production of a particular segment are no longer temporally synchronized).
13. Consult table about muscle (and other) inertia factors.
14. Bring linguistically-determined information about any limits to be imposed on any articulatory variation which is likely to occur.
15. Recruit any additional (antagonistic included) muscles which may be necessary to maintain the limits of articulation variation decided under 14.
16. Set gamma-loop for targets and limits.
17. Send command at appropriate time (notice that this model does not assume the possibility of any low-level sequential triggering of commands).
18. The muscles contract within the specified limits, beginning at the correct time to achieve synchrony of articulator movement associated with a particular segment. Notice that EMG data shows that contraction seems to have finer temporal than amplitude

limits (Tatham and Morton 1968a).

20. Apply the acoustic theory.

21. Output soundwave.

Implementing the above model is well-nigh impossible, for several reasons, principal among which is that there is just not enough data for most of the boxes (even if the boxes themselves are correct). Take, as an obvious example, the temporal compensation and motor-cohesion boxes: that these two phenomena exist seems likely at the present time, as we have shown, but even a simple descriptive statement of their details does not exist yet. For the moment this does not matter. What does matter is attempting to use the model's implications for synthesizing speech even if we have to guess at individual values for any item. Guessing reduces reliability of using the working model for perception research, but it is a way of getting at the details for production research.

Before beginning a description of the synthesis strategy let us recapitulate the most fundamental assumptions of the present model--which (however grossly) would need their respective representations somewhere in the synthesis system.

## Fundamental Assumptions of the Speech Production Model

A The input shall consist of individual segments which shall be extrinsic allophones bearing only notional time marking in the form of simple sequencing.

B The input shall be indexed with boundary symbols, such as: utterance, group, word, morpheme, syllable.

C The input shall be indexed with certain prosodic features, such as: stress (lexical, group, sentence), intonation (possibly only if marked, but suspect all).

D Also input will be (extra-linguistic?) information derived from decisions about the overall rate of utterance.

E Speech production is ultimately reducible to articulatory targets (though whether these are stored as representations of shapes, sounds, muscle commands, etc., is unknown).

F These targets are constant within individual for a particular language (irrespective of final output rate, co-articulation, segment position within the syllabie, etc.).

G The hypothesis is adhered to for the moment that a-linguistic motor control dominates the syllabification of segment sequences at the periphery.

H Rate of utterance does not dominate the programming of targets but merely provides a factor which will enhance the effects of system inertia.

I A function of the motor control system is to stagger (negatively or positively) individual muscle commands to achieve desired articulatory movement at the correct time--the theoretical standpoint is taken that it is not until this late time that staggering occurs. Staggering is computed according to lookup table.

J A lookup table containing inertia factors reacts with command staggering and antagonistic systems, together with psychological/

perceptual information about the crowding status of the perceptual space, to compute limits of co-articulation and variation (over- and under-shoot) which may be permitted.

K Some kind of buffer is required at the point of staggering, serving two functions: (a) it permits successive passes of data for re-arrangement for staggering and (b) permits 'hold' facilities for output to final peripheral mechanisms of the motor-command signals.

L This model (despite Wickelgren 1969) holds that gamma-loops and any similar mechanisms are used for two functions: (a) to 'set' limits and (b) to hold them; it further holds that

M gamma-loop systems are used to provide information about the prior state of the muscles which results in a left-to-right effect observed in the final output; and further

N that fast conducting neurons will permit command signals to arrive at a muscle *during* the previous segment, thus generating the right-to-left effects observed in EMG and other data (MacNeilage and Declerk 1968).

O A composite signal arrives at the muscle which is a temporally-governed transform of the original extrinsic allophone lookup table target values. This signal embodies: (a) the target value (re-computed); (b) a temporal element (which may just be a re-issuing of the same command for a given period of time); (c) limiting factors to govern phenomena associated with any succumbing to inertia.

P It is therefore held as a theoretical tenet that at the final stages of articulation, *universal* (that is: always operating within a particular speaker under normal conditions and comparable with similar effects in other speakers) constraints apply to the transformation of the input signal to the muscles to produce the output configuration. This constraint system is rule-governed (i.e., quite predictable) and 'known' to the system which has used this information to compute the limitations or counteraction measures to be applied.

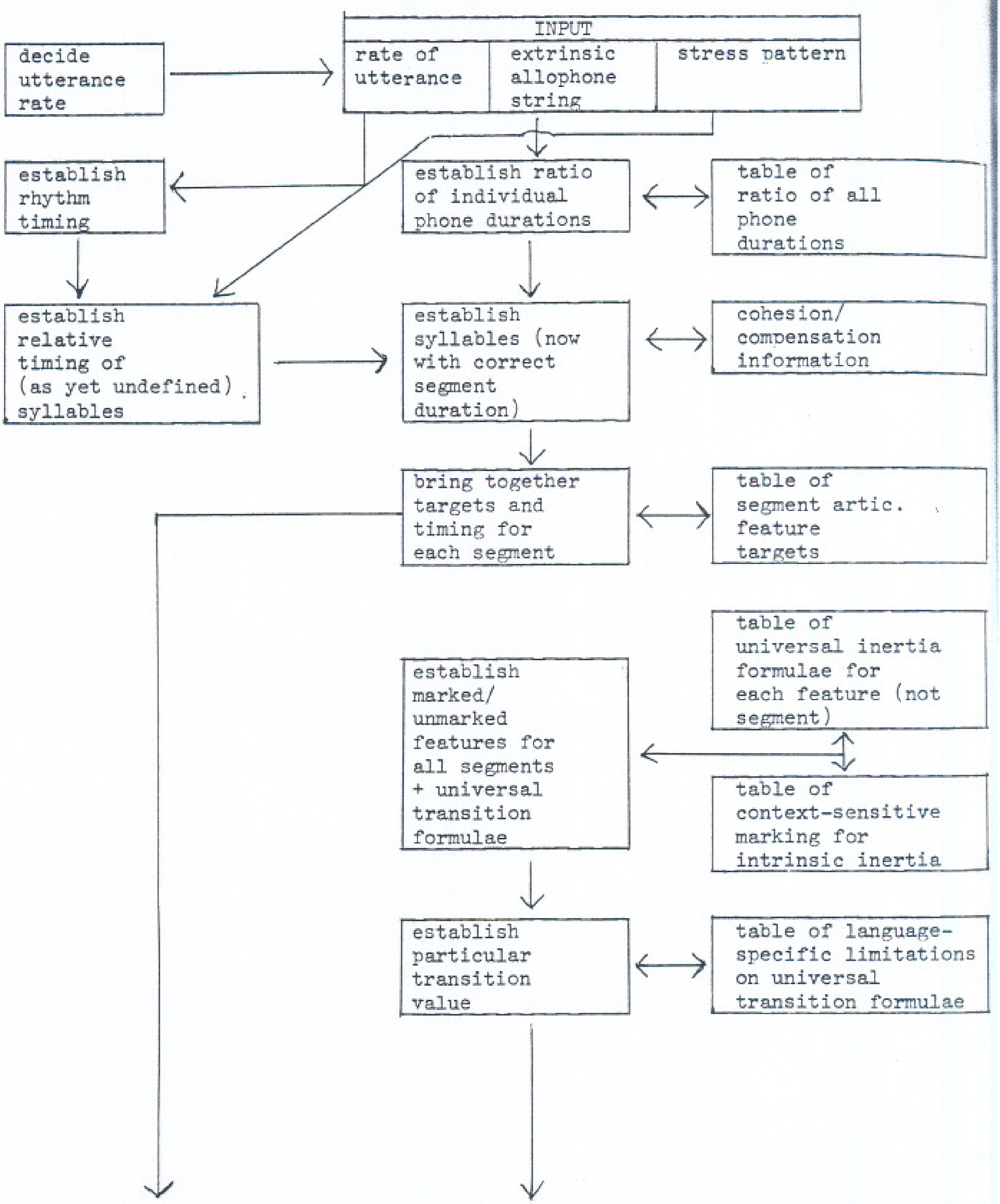
Q Such a postulate of the role of inertia predicts that where constraints were not controllable then no fine differentiation could be required by the linguistic system (a trivial hypothesis, but requiring to be stated).

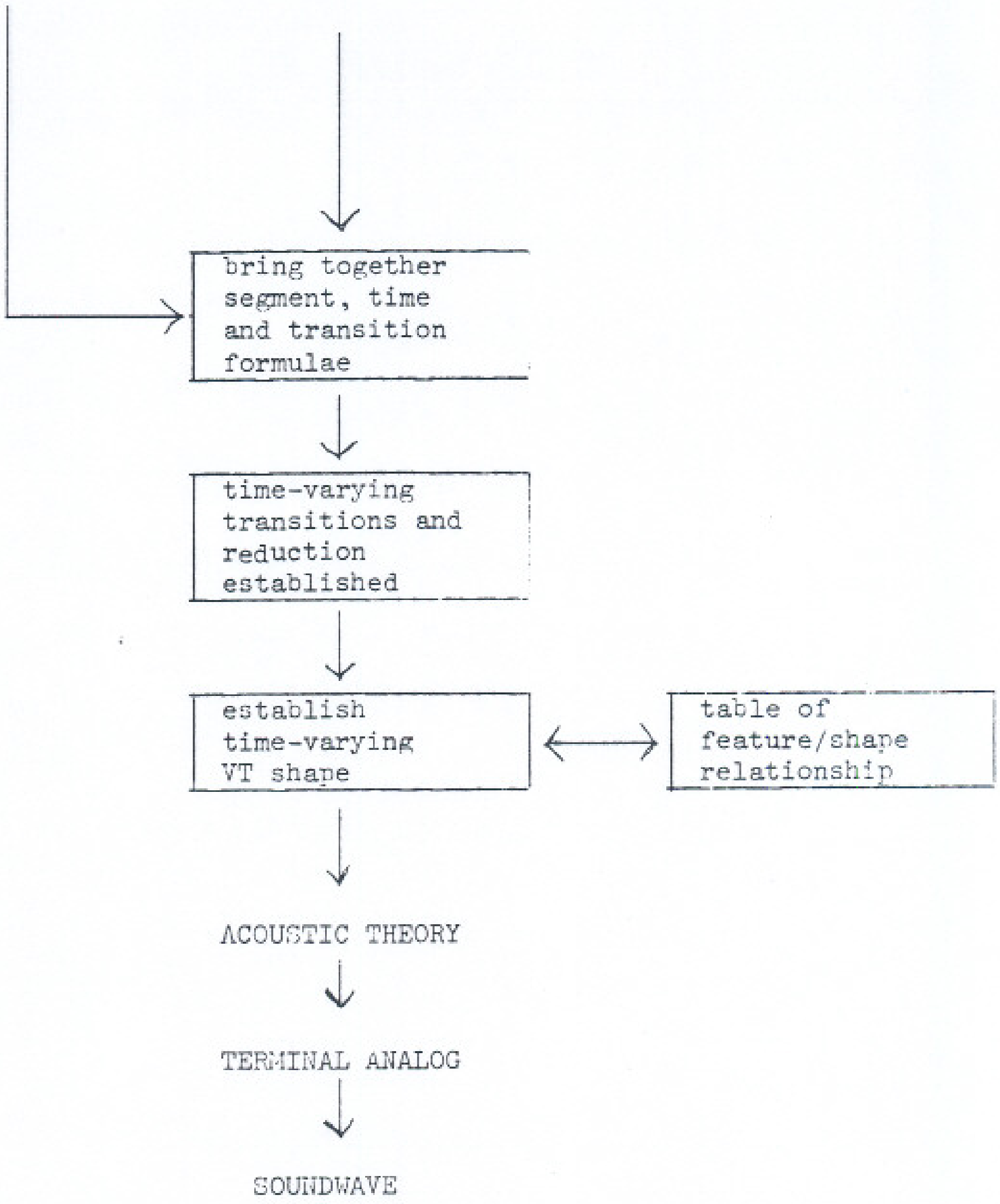
R The model permits variation in successive repetitions of the same utterance in a way that existing speech-synthesis systems do not (we are concerned only with speech synthesis-by-rule). Existing programs (Holmes et al. 1964; Mattingly 1968) store targets and generate allophones in such a way that neither temporal nor target nor transitional output can vary unless the stores are changed.

## 3. The Synthesis Strategy

Fig. 2 is a very tentative suggestion for implementing the above model. Most of the information required for the lookup tables does not exist.

Fig. 2. DIAGRAM OF SYNTHESIS PROCEDURE


INPUT
rate of utterance
extrinsic allophone string
stress pattern
decide utterance rate
establish rhythm timing
establish ratio of individual phone durations
table of ratio of all phone durations
establish relative timing of (as yet undefined) syllables
establish syllables (now with correct segment duration)
cohesion/ compensation information
bring together targets and timing for each segment
table of segment artic. feature targets
table of universal inertia formulae for each feature (not segment)
establish marked/ unmarked features for all segments + universal transition formulae
table of context-sensitive marking for intrinsic inertia
establish particular transition value
table of language-specific limitations on universal transition formulae


bring together segment, time and transition formulae
time-varying transitions and reduction established
establish time-varying VT shape
table of feature/shape relationship
ACOUSTIC THEORY
TERMINAL ANALOG
SOUNDWAVE

Description of the synthesis system

1. Together with the decision to input a certain sequence of extrinsic allophones an overall rate for the utterance is decided; this could take the form 'fast (+), standard (0), slow (-)'.
2. A rhythm is established which might take the form of an actual time for duration between tonic-stressed elements.
3. Information from the input about the stress pattern establishes the placement of stressed and unstressed elements within the established rhythm.
4. By reference to a table of the relative durations of all phones in the language the relative durations of the actual input allophones in sequence are established.
5. (3) and (4) are combined to establish the actual timing of each element (segment) by including information about cohesion and compensation.
6. Reference is made to a table of target values for each feature for each of the segments in the utterance and this information is brought together with the timing information already established.
7. A table of universal inertia formulae associated with each articulatory feature and a table of marking values for these inertia formulae dependent on segmental context are brought together to establish which features of the utterance segments are marked or unmarked for transition and what the transition formulae are for these features on a universal basis.
8. By lookup table of the language-specific limitations to be applied to these transition formulae, particular formulae are substituted for the output of (7).
9. Feature targets, segment timing and specific transition formulae are brought together.
10. Transitions and reduction, etc., are computed.
11. VT shape is established by means of a lookup table related to the output of 10.
12. Acoustic theory is applied.
13. Conversion of the output of the acoustic theory to TA parameters.
14. Operation of TA to output (5) soundwave.

Notice that prior input information for storage purposes (lookup) is required for:

(a) general ratio of phone durations
(b) cohesion/compensation information
(c) articulatory-feature targets for all segments
(d) universal inertia formulae for each feature
(e) context-sensitive inertia information
(f) language-specific limitations of inertia values
(g) feature/shape relationship.

Hypothesis: (a), (f) and (?b) are language specific; the rest are universal.

Utterance-specific input information required:

(a) string of extrinsic allophone segments
(b) rate of utterance

(c) stress information
(?d) intonation

Some of the deficiencies

A The system is tentative for the moment and much of it could not be implemented except on a very ad hoc basis because values for most of the lookup table information are not available.

B In particular no mention has been made of intonation and how this is derived; no mention has been made either of how amplitude control (e.g, for stressed vowels) is derived.

C Quite clearly not all the boxes in the speech production model have been implemented (particularly at the neuro-muscular level).

D Particularly unsatisfactory is the way segment, timing and transition formulae (9) are brought together in one big obscure computation.

It is hoped that this is the beginning of a speech-synthesis-by-rule system which will render transparent some of the stages in the speech production process.

References:

Boomer, D. S. and J. D. M. Laver (1967) "Slips of the Tongue," Work in Progress No. 1, U. of Edinburgh Linguistics Department.

Broadbent, D. E. and P. Ladefoged (1960) "Voice judgments and adaptation level," Proc. Royal Soc. B. Vol. 151.

Cooper, F. S. (1965) "Research techniques and Instrumentation: EMG," Proc. Conference: Communicative Problems in Cleft Palate: ASHA Report No. 1.

Drachman, G. (1970) "Rules in the Speech Tract," Papers from the Sixth Regional Meeting of the Chicago Linguistic Society, University of Chicago.

Fant, C. G. (1960) Acoustic Theory of Speech Production, Mouton: The Hague.

Fant, C. G. (1965) "Formants and Cavities," Proc. 5th International Congress of Phonetic Sciences, Münster, 1964, Karger: Basel/ New York.

Flanagan, J. L. (1965) Speech Analysis, Synthesis and Perception. Springer Verlag: Berlin.

Holmes, J. N., I. G. Mattingly and J. N. Shearme (1964) "Speech Synthesis by Rule," Language and Speech 7.

Lawrence, W (1953) "The Synthesis of Speech from Signals which Have a Low Information Rate," Communication Theory, ed. by W. Jackson: New York/London.

Lehiste, Ilse (1970) "Temporal Organization of Spoken Language," Working Papers in Linguistics No. 4, Computer and Information Sciences Research Center, Ohio State University.

Lieberman, A. M., P. Delattre and F. S. Cooper (1954) "The Role of Consonant-vowel Transitions in the Perception of the Stop and Nasal Consonants." Psych. Monograph 68.

Matthews, P. B. C. (1964) "Muscle Spindles and Their Motor Control," Physiol. Review 44.

Mattingly, I. G. (1968) "Synthesis by Rule of General American English," Supplement to Status Report on Speech Research: Haskins Labs: New York.

MacNeilage, P. F. and J. L. Declerk (1968), "On the Motor Control of Coarticulation in CVC Monosyllables." Haskins Labs, SR-12: New York.

Öhman, S. E. G. (1964) "Numerical Model for Coarticulation Using a Computer-simulated Vocal Tract." JASA 36.

Öhman, S. E. (1966) "Co-articulation in VCV Utterances: Spectrographic Measurements," JASA 39.

Öhman, S. E. G. (1967) "Peripheral Motor Commands in Labial Articulation," STL-QPSR 4, 1967. RIT: Stockholm.

Tatham, M. A. A. (1969a) "The Control of Muscles in Speech," Occasional Papers No. 3, U. of Essex Language Centre.

Tatham, M. A. A. (1969b) "Classifying Allophones," Occasional Papers No. 3, U. of Essex Language Centre; also to appear in Language and Speech 1970.

Tatham, M. A. A. (1970) "Speech Synthesis: A Critical Review of the State of the Art," *Int. Journal Man-Machine Studies* Vol. 2.

Tatham, M. A. A. and Katherine Morton (1968) "Further Electromyography Data Towards a Model of Speech Production," *Occasional Papers No. 1*, Univ. of Essex Language Centre.

Werner, Edwenna and M. Haggard (1969) "Articulatory Synthesis by Rule," *Speech Synthesis and Perception Progress Report No. 1*. Psychological Laboratory, U. of Cambridge.

Wickelgren, W. A. "Context-sensitive Coding, Associative Memory and Serial Order in (Speech) Behavior." *Psychological Review* Vol. 79.1.

# A Model of Speech Perception by Humans

L. V. Bondarko, N. G. Zagorujko, V. A. Koževnikov, A. P. Molčanov,

and L. A. Čistovič

Translated from Russian by Ilse Lehiste *

*Sponsored in part by the National Science Foundation through Grant GN-534.1 from the Office of Science Information Service to the Computer and Information Science Research Center, The Ohio State University.

# A Model of Speech Perception by Humans

L. V. Bondarko, N. G. Zagorujko, V. A. Koževnikov, A. P. Molčanov,
and L. A. Čistovič

Translated from Russian by Ilse Lehiste

## Foreword

This book sets forth some results of investigations in the areas of psychology, physiology, and experimental phonetics, directed towards the elucidation of the mechanism of speech perception by humans. On the basis of these data and the application of methods of the theories of complex systems and pattern recognition, a plausible model of speech perception by humans is presented.

The work may be of interest to specialists working in the area of the automatic recognition of speech signals: mathematicians, engineers, physiologists, psychologists, and linguists.

## 1. Introduction

The authors of this work are united in the conviction that the elaboration of a model for speech perception by humans coincides in practice with the elaboration of a system of automatic recognition of a sufficiently large set of speech events.

It is not necessary (and, for the time being, not possible) to demand complete structural isomorphism between the human speech perception system and the system of automatic recognition of speech signals. One can, however, hope for a functional resemblance between the model and the original.

In the process of developing the model, it is unavoidable that questions arise which are inaccessible (or accessible with difficulty) to direct experimental investigation. Insufficient information is then supplemented by guesses and assumptions. The first natural test which these assumptions must meet consists in the requirement that the model which has been set up using these assumptions must be efficient. This, of course, cannot be established before the model is converted into a technical construct or a machine algorithm.

Is it impossible to solve the problem of constructing a model of speech perception in a purely formal way?

For example, let us try to look at the procedure of speech recognition from the point of view of the theory of complex systems [1]. If one does not demand structural isomorphism between natural structures and those to be designed, then it is possible to assume an infinite number of variants of the automatic speech recognition device. Optimal will be the automaton that will recognize a given lexicon with the required reliability $P_o$ at a minimal cost R. Obviously R will be a function of the cost of memory elements (short-term as well as long-term memory) and other elements which enter the construction.

Hardly any speech researcher believes at the moment that a sufficiently large set of words can be recognized immediately from current parameter values of the speech signal. It is the experience of many laboratories that this method of approach is justified only when the vocabulary does not exceed 20-25 words.

For more complex tasks, multi-stage hierarchical structures of recognition devices are usually proposed. A general block diagram of a multi-stage recognition device is presented on Fig. 1.

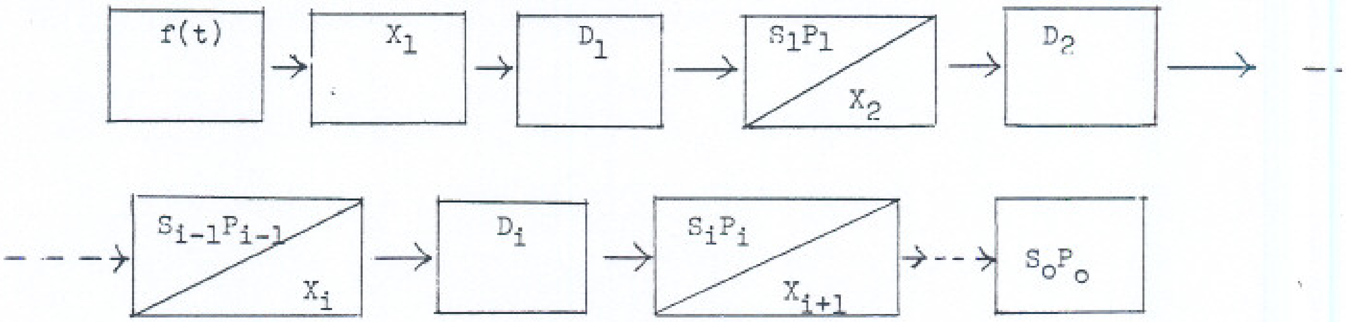


Figure 1

Here $X_i$ constitutes the description of the signal at the input to the i-th level of perception. The classifier $D_i$, with the help of certain rules ("decision functions") makes the decision whether an unknown vector-realization $X_i$ belongs to this or that element of the alphabet $S_i$ with a reliability $P_i$. The sequence of elements $S_i$ constitutes the selection space $X_{i+1}$ of the next (i + 1th) level of the recognition device.

The diagram given on Fig. 1 is a general scheme which may serve as a skeleton model for the analysis of existing artificial as well as natural hierarchical recognition devices. The analysis is reduced to determination of the number of levels and the structural elements of each level X, D, S, and P, and the nature of the interaction between these elements.

Externally given requirements for the planner are usually the speech signal f(t) at the input, the lexicon $S_o$, and the reliability $P_o$ at the output of the system. All intermediate blocks may be selected arbitrarily.

First of all, it is indispensable to implement the transition from f(t) to a more compact and at the same time sufficiently informative description $X_1$. It is possible to try several, for example "$k_1$", variants of description. Certain short speech elements $S_1$ will be recognized according to this description with a reliability $P_1$, after them--elements $S_2$ in their turn, etc. Classifiers may have a different structure at every level. It would be desirable to examine several ("r") versions of algorithms for decision-making.

There are no formal limitations on the intermediate alphabets $S_i$. At every level it is possible to examine the suitability of "q" variants of the alphabet. If the number of stages equals m, then N different variants of recognition systems are subject to examination, whereby

$$N = k_1 \times r_1 \times q_1 \times r_2 \times q_2 \ldots . r_m = k_1 \times r_m \prod_{i=1}^{m-1} r_i q_i .$$

For the sake of simplicity, let us assume that $k_1 = r_i = q_i = n$. Then $N = n^{2m}$, and when $m = 4$, and $n = 5$, $N = 360{,}000$. It is clear that the examination of so many variants is practically impossible, especially considering the fact that each variant represents a very cumbersome task. The most expedient is the following 'biotic' approach: for a first approximation to the optimal schema, the variant should be selected which incorporates all reliably known facts concerning the physiology and psychology of speech perception. Later on it might be attempted to find the best approximation to the optimum in the neighborhood of the point represented by this variant.

As was noted above, the general operating criterion for testing the quality of systems being projected is the summary complexity (cost) of the system R, complying with the limitations $f(t)$, $S_o$ and $P_o$. It is possible to determine the complexity of each stage separately. For example, the complexity of the classifier $D_i$ is a function of such quantities as the extent of the selection space (i.e. the number of elements in the lexicon of the preceding stage $S_{i-1}$), the number of elements to be recognized $S_i$ and the type of the decision function $D_i$. The selection of $D_i$ depends in turn on the required reliability of recognition $P_i$ and on the level of reliability $P_{i-1}$ with which the elements $S_{i-1}$ had been recognized. Therefore

$$R_i = f(S_{i-1}, P_{i-1}, D_i, S_i, P_i).$$

Unfortunately the shape of this function is unknown: it is not known how the reliability of recognition Pi-1 is related to $P_i$, that is, what mistakes in the recognition of elements $S_i$ are induced by mistakes made in the recognition of elements at the preceding level $S_{i-1}$. At any rate, the form of this dependence must be determined experimentally (that is, it is necessary to construct and test a concrete automatic machine). This constitutes the basic reason why up to now it has not been possible to optimize a hierarchical structure by formal methods of optimization of the type employed in linear or dynamic programming. For setting up a more detailed schema of the automatic machine at each level it is therefore indispensable to return to known facts about the perception of speech signals.

Thus both the investigators of speech perception (phoneticians, psychologists, physiologists) and the specialists in the automatic recognition of speech (mathematicians and engineers) are now equally interested in setting up a first version of a speech perception model which would incorporate in an unequivocal manner positively known facts about speech perception by humans. The present paper reflects the first stage of our collective efforts in this direction.

In the beginning of the paper (§2) facts and assumptions are presented regarding the structure of a model of speech perception. This is followed by an exposition of some elements of this model (§§3, 4, 5). The paper concludes with a schema and description (§6) of a plausible (from our point of view) version of a model of the recognition of speech signals.

## 2. Structure of the Speech Perception Model

As soon as we define the final result of speech perception as understanding the meaning of the communication, and demand that it should be possible to understand sentences that have never been heard before, it becomes obvious that the process of perception must be hierarchically organized.

In order to understand the meaning of a sentence it is indispensable to have at one's disposal a description of the syntactic structure of the sentence. In order to carry through a syntactic analysis, it is indispensable to have the sentence first divided into words, and to have assigned to each word its lexical and grammatical characteristics. In order to analyze a word, it is preliminarily necessary to have at one's disposal its phonemic or near-phonemic description. Finally, in order to transform a speech signal into a sequence of phonemes, it is first of all necessary to distinguish in that signal those acoustic features that differentiate phonemes from each other. Distinguishing among acoustic features presupposes an earlier time-frequency analysis of the stimulus.

Specialists engaged in the automatic analysis of texts consider it to be sufficiently well established that the transformation of the alphabetic form of a sentence into a description of its meaning must consist of three successive stages, represented on Fig. 2 to the right of the dashed line [2, 3, 4]. It is obvious that the same three stages must also be present in the analysis of spoken language. Furthermore, the process of the perception of spoken language must include at least two additional preliminary stages of transformation [5, 6, 7, 8].

The first of these stages is the auditory analysis of the speech stimulus. As a result of the operation of this stage, the stimulus is described in terms of acoustic (auditory) features. Logically it is to be expected that the set of features which the auditory system distinguishes in the signal is quite large and is intended for the totality of acoustic signals with which the organism has to deal. It is probable that only a part of these auditory features will turn out to be useful for speech recognition.

The next stage in perception is the phonetic interpretation of the stimulus. The description produced at the output of this block must be already sufficiently abstract and applicable to either an acoustic or an articulatory representation of the speech event. Such an abstract description might be given in terms of, say, phonemes or distinctive features [9].

## 2.1. The hierarchical model and the possibility of its realization with the help of simple automata.

From the point of view of the theory of complex systems, one of the advantages of hierarchical structure is the fact that each block can be relatively simple ("cheap"), can make do with small amounts of short-term and long-term memory and with a limited number of operations in decision-making.

This is connected with the fact that each preceding block serves as an information filter with respect to the following block, decreasing the dimensions of the signal and bringing it closer to a form that is more convenient for further processing.

Let us imagine that classifier $D_i$ has been allotted a limited number of cells of short-term memory and a limited number of operations. Then it must inevitably be simple, for example linear, and must operate within a restricted space, and the level i - 1 must output such elements $S_{i-1}$ that short sentences thereof can be recognized at the i-th level with the help of linear decision functions. The possibility of replacing a complex decision function with a sequence of simple (linear) classifiers is demonstrated in reference [10]. If the summary complexity (cost) of a multi-stage system with an identical reliability ($P_o$) in the recognition of elements $S_o$ turns out to be less than the complexity (cost) of a single-stage system, then a re-coding may be considered justified.

To what an extent are these arguments in favor of hierarchical structure supported by facts about the auditory analysis mechanism?

It is considered to be sufficiently well established at the present time that the capacity of the human short-term memory is very limited [11, 12]. This is revealed by data concerning the retention of speech or speech-like stimuli. Thus it has been shown that a sequence consisting of only three vowels or pure tones is remembered as a sequence of decisions about stimuli and not as a sequence of auditory descriptions of stimuli [13]. This makes it possible to believe that the automaton carrying out a phonemic interpretation of the stimuli must perforce work with auditory descriptions of very short segments of the speech train, certainly shorter than the average duration of a word. It is known that the length of a sequence of nonsense words which a human can remember does not exceed 7 - 10 syllables [11, 14]. This obviously characterizes the dimensions of the "temporal window" through which the utterance is "seen" by the automaton performing the morphological analysis of the word. The sequence of grammatically and semantically unconnected words which a human can reproduce after one hearing is likewise very limited [12].

Finally, it has been demonstrated that it is easier to recall the meaning of a sentence than the complete sequence of words constituting the sentence [15]. The adduced data allow one to assume that from the point of view of the total cost of short-term memory at all levels, the hierarchical system of speech recognition must prove sufficiently economical.

Let us proceed further. It is known that the complexity of the classifier depends to a very high degree on the number of patterns to be recognized. Even if we assume that we should succeed in using alphabets of small dimensions at intermediate stages, nevertheless at the last stage the alphabet of objects to be recognized cannot be smaller than, say, the number of words in the lexicon $S_o$. Would it not be possible to recognize a word without comparing its complete description with every standard item contained in the lexicon $S_o$?

In reference [16] an algorithm is described of a step-by-step reduction of the lexicon in the recognition process, which is based on the method of "crossing out" proposed by L. Čistovič. In this method, a feature is selected (the first one that occurs or the first in importance among a number of simultaneously occurring features), and all words that lack this feature or this particular meaning of the feature are crossed out from the initial lexicon $S_o$. Thus the lexicon is sharply reduced. The same operation is performed with other features. The task becomes simpler at every step. At a specified stage, the algorithm proceeds to a comparison of the word with the standard forms of the remaining words in the lexicon, in the complete, multi-dimensional description space, with the help of any chosen decision function. This "combined" algorithm enables one to reduce the decision-making time by several orders of magnitude. There exist reasons to assume that humans follow an analogous procedure. It would be important to find out how concretely it is realized at every hierarchical level of human perception.

An analogous role--reduction of the initial lexicon on the basis of incomplete preliminary information--is probably played by the phenomenon called "psychological setting"--the increase of the a priori probability of certain hypotheses as compared with others. The algorithmic model of this procedure differs hardly at all from the "crossing out" procedure and is possibly realized in living systems by means of a general physiological mechanism.

The reduction of hypotheses and numbers of variants apparently plays an important role at every level of the system, which makes it possible to employ economical classifiers.

A study of the nature of decision functions used by humans has shown that in the process of making a decision in a multi-dimensional space of features, they employ hyperplanes parallel to the planes of coordinates, i.e., the simplest type of linear decision functions [17]. An analogous result was obtained in experiments dealing directly with the perception of speech signals [18]. Consequently, experimental facts support the arguments (the small capacity of short-term memory and the simplicity of classifiers) used to justify the advisability of the hierarchical structure of recognition.

## 2.2. The hierarchical model and the reliability of recognition.

It is natural that at every level of the hierarchical system information losses must take place, which seemingly makes such a system less effective with respect to the reliability of recognition in comparison with a single-stage system. This question has been repeatedly discussed in the literature in connection with the problem of 'decision units' in speech perception [19]. There exists a large amount of trustworthy experimental data, which demonstrate that in the interpretation of a speech stimulus one relies not only on its acoustic properties, but employs also information concerning phonological and syntactic rules, frequency characteristics of the lexicon, and semantic rules (cf. the survey in ref. 5). It can be concluded from this that one does not make decisions about individual phonemes in the stream of speech and that the units with which one operates correspond to words or even larger segments [19].

If we should interpret this result to mean that acoustic images of whole sentences must exist in human brains, we would arrive at complete absurdity, since we would be forced to assume the presence of images of sentences that have never yet been heard. A reasonable explanation of this result might be the following: if the information at the input to a given classifier proves insufficient, the classifier outputs several possible interpretations of the input signal indicating their a posteriori probabilities, and the final decision in the sense of the selection of one definite alternative may be postponed from stage to stage all the way up to the last one--the recognition of the meaning of the utterance.

The presented ideas correspond to the conclusions drawn by Galunov [20] on the basis of an experimental investigation of the perception of speech in noise. The author arrived at the conclusion that humans carry out a continuous re-coding of the speech stream into phonemes, but make final decision after the elapse of sufficiently long segments.

There is no doubt that the stability of spoken communication among humans in spite of interference is based on the use of redundancy. Voloshin worked out an algorithm for increasing the reliability of recognition at the expense of the redundancy of the signal [21]. Experimental testing of the algorithm showed that it is indeed possible to build reliably functioning devices for the recognition of oral commands on the basis of phonemes that are recognized with a low reliability.

The complexity (or the reliability) of the automatic recognition device depends strongly on the nature of the distribution of the totality of objects to be recognized in the selection space. Usually, given the recognition reliability $P_i$ of elements of the alphabet $S_i$, the complexity of the classifier $D_i$ increases with a greater dispersion of those elements $S_i$ in the space $X_i$. An increase in the dispersion of speech signals usually accompanies an increase in the number of speakers who participate in the experiment.

It would be possible to decrease the dispersion, if one could successfully limit oneself to working with standard forms produced by one speaker, adapting them to the particular characteristics of any other speaker. The possibility of such a procedure was examined in [22]. It turned out that the quality of recognition is in fact substantially increased when a standard is provided for the speaker.

The use of a hierarchical recognition system may allow one to make use of the most varied kinds of information about the speaker. starting from the acoustic peculiarities of his pronunciation and ending with the sphere of concepts with which he operates. One possible mechanism could be changing the a priori probabilities of output units in the lexicons of classifiers. It is possible to think that some kind of elementary adaptation to the speaker is already incorporated at the level at which auditory features are isolated. Thus the phoneme boundary in the space constituted by the first two formants of isolated synthetic vowels depends on the frequency of the fundamental tone and on the frequency of the third formant [23].. The reliability of recognition is also obviously increased through adaptation with respect to tempo, speech loudness, acoustic characteristics of the room, etc.

## 2.3. Special characteristics of the proposed model

It follows from everything said above that a complete model of speech perception must include such higher stages of information processing that are currently being investigated by specialists in machine translation. The realization of such a complete model in the form of automatic algorithms is hardly possible in the near future. At the same time it is obvious that partial models, describing the transformation of information at separate stages of the chain, should preferably be worked out in such a way that they could later be easily inserted into one general model. For that purpose it is indispensable that the output signals of models of lower levels be identical with input signals to models of higher levels.

At the present time specialists in machine translation work with written texts, and input signals for their algorithms are written words, i.e., strings of letters separated by spaces. In oral speech there are usually no pauses between words, and the problem of determining what is a word appears to be sufficiently complex in itself. Besides phonemic information there is also prosodic information which likewise must be transmitted in a transformed shape of some kind to the input of the block that carries out the syntactic analysis. This compels us to assume that the model for morphological analysis (block 3 on Fig. 2) must be worked out specially in conformity with requirements for oral speech, and that this is a task for the joint efforts of specialists in automatic speech recognition and specialists in machine translation.

Correspondingly, we shall formulate the task of the present investigation as producing a model of the chain of transformations that ensure the transition from an acoustic speech signal to its description in terms of a sequence of words, in which every word in its turn is described in terms of the set of its lexical and grammatical features. This corresponds to the first three blocks on the schema presented on Fig. 2. Besides that, a word must be assigned at the output certain supplementary prosodic characteristics (this question is not at all clear yet and requires special investigation).

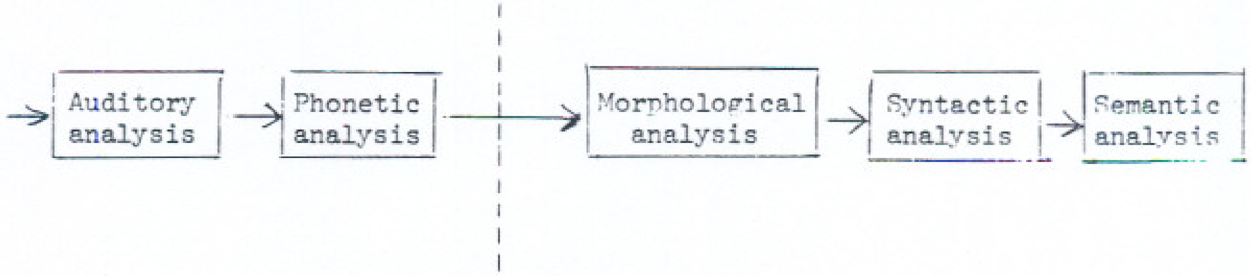


Figure 2.

In the beginning stages the size of the lexicon of the recognition device will rationally be limited to two-three thousand words.

Below we will attempt to formulate some considerations regarding the structure of the first two blocks of the model. Without touching the structure of the third block, we will at this time only define the character of its input signals. It is possible to consider them as sequences consisting of no more than ten syllables each. A syllable will be defined as an element ending in a vowel and containing no more than one vowel. Each syllable is described in terms of phonemes (which, however, may be incompletely recognized). In addition, each syllable is provided with supplementary information characterizing its stress and intonation (the number of degrees and the form of description of these features has not yet been determined).

It has been shown in the work of Lisenko [24] that on the basis of such input information it is, in principle, possible to arrive at a segmentation into words and carry through their morphological analysis.

A specification of the input signals for the third block is indispensable within the framework of this study, since it determines the requirements which must be imposed on the phonetic interpretation block (block 2 on the schema presented on Fig. 2). Let us now turn to facts which are known from electrophysiology of hearing and from psychoacoustics.

## 3. The processing of speech signals by the auditory organ.

The initial analysis of signals by the auditory organ takes place in the cochlea. Because of the non-uniform structure of the basilar membrane, the transmission of energy from the signal to its various points is realized with dispersion in frequency and time.

In practice, a spectral analysis of the signal takes place in the cochlea according to the transmission functions of the basilar membrane. An approximation of those to Bekesy's data was carried through by Flanagan [25]. An examination of the results of the approximation shows that the representation of the stimulus at a given point of the cochlea is connected with a definite frequency selection and temporal delay of the signal.

It is known from psychoacoustic experiments that a decision about the stimulus is taken with regard to its development beyond the so-called critical time of analysis, consisting of approximately 150-200 milliseconds. Taking into account what has been said, the representation of the signal at the level of the cochlea may be described by means of a contour in the frequency-time-energy space. A given contour reflects the stimulus which immediately brings about the excitation of neuron endings found in the organ of Corti. However, the transmission of this excitation to successive sections of the neural net proceeds differently for different elements of the contour causing the excitation. Psychoacoustic experiments demonstrate that a human listener identifies sound signals as similar if they differ only in amplitude. Invariance with regard to amplitude (loudness normalization) is evidently connected with the transmission function involved in the transmission of the excitation into the neural net. The frequency of impulsation (or the probability that a response occurs after a given time segment) of a peripheral neuron is connected with the intensity of the stimulus that acts upon the corresponding receptor with the logarithmic dependency relation

$$p\,(x) = \log\frac{1}{t}\int_{o}^{t} f^{2}_{t}\,(t')\,dt' \sim \log\,(E + P_{o}); \qquad (1)$$

where $p(x)$ - the probability that an impulse will occur in response to the stimulus,
$P_o$ - the quantity accounting for the spontaneous activity of the neuron,
$t$ - integration time,
$f_t(t)$ - the size of the instantaneous value of the stimulus (only its positive part is taken into account).

If the part standing under the logarithm sign in equation (1) is taken to be connected with the average energy of the stimulus, then the response reaction of the neuron at the periphery of the auditory system will be proportional to the logarithm of the average energy of the stimulus E for the time t. The working rule of the elementary structure that isolates any given feature which is dependent upon the spatial distribution of the energy of the stimulus along the basilar membrane and independent of changes in intensity, may be formulated in the following way:

$$A_i = p(x_1) - p(x_2) = \log \left( \frac{E_1(x)}{E_2(x)} + E_o \right), \qquad (2)$$

where $x_1$ and $x_2$ - two arbitrary points on the basilar membrane,
$E_o$ - a constant taking into account spontaneous activity.

It is possible to write an analogous expression for describing the amplitude changes of the stimulus with respect to time. A complete description of the shape of the stimulus, invariant with regard to its intensity, may be obtained if the values represented in equation (2) are formulated for all distinctive points of the excitation contour. It seems that the just presented normalization process of the description of the signal with respect to intensity takes place at the outermost periphery of the neural net and constitutes a part of the mechanisms for isolating the most distinctive points. It is of great interest to examine electrophysiological data as to which elements of the signal are observed to produce the most clear-cut reactions of the neurons in the various sections of the neural part of the organ of hearing.

At the present time, abundant data are available regarding the responses of single neurons, starting from bipolar cells and ending with the cortical part of hearing. It is difficult to say to what an extent the characteristics of these responses are connected with the results of psychoacoustic experiments. Nevertheless, the fact that there exist selective responses of neurons to signals of a particular form constitutes evidence that at least at the periphery of the system of hearing, a representation of the signal is formed which is based on isolating its specific features.

In reference [28] it is demonstrated that there exist two groups of neurons in the cochlear nucleus which can be clearly separated according to the nature of their responses, conditionally called tonal and phasal. The former are characterized by a substantial sharpening of the frequency-threshold curves when the duration of tonal emission is increased from 2 to 100 msec, by a significant dependence of the latency period on the intensity of the signal, by the preservation of the response during the whole length of the stimulus, and also by a clearly expressed temporal summation of the energy of the stimulus. The character of the frequency-threshold curves of the neurons with tonal response itself points to a significantly sharper reaction of the observed neuron to the stimulus at a given frequency than could have been expected on the basis of the frequency characteristics of the inner ear's mechanical system (the so-called sharpening effect). Among the existing hypotheses set up to explain the sharpening effect, that one appears best founded that proposes the existence of some kind of lateral inhibition at the periphery of the neural part of the organ of hearing.

The simplest variant of lateral inhibition, conditioning the sharpening effect, appears to be a scheme for isolating the differences in the intensity of excitation of adjacent elements. For carrying through this operation, the existence of neuronal structure is postulated that reacts to non-uniformity in the distribution of energy in the space of the receptors.

In the limiting case, such a structure could consist of a single element of a neuron, if the response to a stimulus is proportional to the sum of the absolute values of the differences in the influence exerted on its dendrite system by the receptors. If this is so, then the response of the neuron will be proportional, in the limiting case, to the derivative of one or another order of the function that describes the spatial distribution of energy in the studied section of the receptive field.

An examination of the model of such a scheme of excitation allows one to note the following of its properties. The quality of frequency-selective characteristics, extracted when a tonal signal is fed in at the input of the model, can raise by an order of magnitude the quality of analogous characteristics of input filters (in this case, the frequency characteristics of the basilar membrane). The angle of the slope of the frequency-selective characteristics may reach a magnitude of the order of hundreds of decibels per octave [26]. If two tonal signals act upon the input of the model simultaneously, of which the second is out of tune relative to the mean frequency, one observes a clearly expressed masking of the first signal. The response of the model is insignificant when a signal with a continuous uniform spectrum is fed in at the input. When tonal signals with varying duration are used as stimuli, the model displays clearly expressed effects of temporal summation: an increase in the duration of the signal is accompanied by a lowering of the threshold of exhaustion and an increase in the quality of the frequency-selective characteristics. The latency period of the response depends strongly on the intensity of the stimulus.

The quoted data show that the model of lateral inhibition in the given formulation possesses the basic properties of neurons with tonal response. The characteristic property of the described lateral inhibition model is the sharp isolation of extremes in the spatial energy distribution of the signal. This allows one to propose that the mechanism for isolating formants in speech signals as starting-point features operates with data about the position of extremes in the continuous spectrum of speech elements.

The neurons which give a so-called phasal response to stimuli are characterized by the independence of the response latency period of the signal intensity, by a small dependence of the response threshold and the sharpness of frequency-threshold curves on the duration of the stimulus, and by a significant dependence of the response on the steepness of the onset of the signal. The phasal response usually consists of one or a few bursts, immediately following the onset of the stimulus. Neurons with phasal response constitute about 20% of all investigated elements in the cochlear nucleus. Besides that, their procentual share increases in the higher sections of the neural net of the auditory system (in the inferior collicula etc.).

The described properties of neural structures with phasal response make it possible to assume that they play the role of determining the moments at which a change takes place in the energy of the signal, concentrated in one or another frequency region. One may assume that a temporal segmentation of the uninterrupted stream of speech is worked out in higher sections of the auditory system on the basis of signals which have been received from phasal-type neurons. Apparently there exist in the neural net of the auditory system isolators of significantly more complex characteristics of stimuli, which describe in detail how their energy changes with respect to time as well as to frequency. The same isolators accomplish the quantitative evaluation of the shape of the perceived signal.

In reference [28] it is shown that in the inferior collicula of rats there are neurons which respond with a group of impulses to a short signal (of the order of 1 msec) and do not respond at all to longer-lasting stimuli (longer than 10 msec).

In the auditory part of the cortex neurons have been isolated that respond only to stimuli whose frequency changes in one or the other direction [30].

The neurons of a given group can be differentiated into three groups according to the nature of their response to a frequency-modulated signal.

The first group is constituted by elements that react to a change in the frequency of a tone in the direction of a higher frequency.

The neurons of the second group respond only when the frequency of a tonal stimulus is decreased (with a definite speed!). Finally, neurons of the third group react to a change in the frequency of a tone, if the direction of this change leads to a closer approximation of the frequency of the tonal signal to the characteristic frequency of the given neuron. A multitude of data, accumulated in investigations of the visual analyzer, point to the existence of structures in its neural part that isolate from external stimuli such elements like contours, angles and more complex configurations [27]. A special significance have those cases in which neurons in the cortex respond only to a spatial movement of a certain kind of signal. It is interesting that the underlining of contours in drawings is realized at the periphery of the visual analyzer of a frog owing to the interaction of the processes of excitation and inhibition, which evolve in time according to different laws [27]. This points to the subtlety and complexity of the structures which accomplish the isolation of informative elements in images. In view of the proposed communality of the basic principles of the processing of information in different sensory systems of the organism, it is useful to consider the possibility that there may exist neural structures in the system of hearing that react to the same kinds of features of signals as are reacted to in the visual analyzer.

Finally some hypotheses should be mentioned that propose the existence of neural structures in the auditory system that isolate certain sequences of occurrences of maxima or sharp decreases of energy in the stimulus at different frequencies and at different moments in time. The existence of such structures might explain the selective reaction of living creatures to certain sounds with a complex spectral and temporal structure. According to current views, neural structures that react selectively to complex signals are continuously being formed in the neural net in the process of developing conditioned reflexes. However, it is necessary to distinguish structures that come into being as a result of training in the perception of new signals from those that are genetically consolidated.

It is proposed at the present time that neural structures which appear in the process of learning are to a certain degree connected with the mechanism of memorizing, and that they participate as a part of this mechanism in the decision-making process regarding the recognition of signals in the extreme stages of analysis. On the other hand, inherited neural structures participate in the beginning stages of the perception of stimuli; they isolate, i.e., react more sharply to those elements that are most characteristic for the whole broad class of signals which is taken in by living creatures of a given species in the processes of vital activity.

The proposition that a number of neural structures that isolate complex elements of stimuli are transmitted genetically was proven for the visual analyzer by direct physiological experiments.

It may be concluded from what has been said above that the representation of an external stimulus in the reaction of an ensemble of neurons that constitute the neural net may be considered passive only at the lowest level (the level of cochlear receptors). Later on, a reaction to the stimulus is formed from responses of those neural structures that isolate certain elements of the signal from its complete description. The presence of a given element in the perceived signal is indicated by the response of a neuron or a group of neurons at the terminal point of a specialized neural structure. Thus it is assumed that the isolation of certain physical features of the signals is accomplished already at the periphery of the neural net. The admission of such features is determined by the availability of corresponding neural structures. The latter are evidently developed in the process of evolution.

This reduced representation of the stimulus is transmitted to the succeeding level of the neural system, where, on the basis of identified features, further operations are carried through in the classification and recognition of stimuli. It may be assumed that not all detected features are used at the next level of analysis, but only some of them (more precisely--the minimal number needed for making a classificatory decision with the necessary degree of reliability). The complete set of features is necessary only for solving the most difficult tasks of classification, when the ensemble of stimuli to be recognized is sufficiently large. For ordinary tasks, the use of only some of the isolated features appears to be sufficient. Therefore it is natural to suppose that the selection of necessary features takes place during the process of solving the problem of classification at a relatively higher level of analysis. Consequently there must exist the possibility of transmitting various combinations of features up to the moment at which a reliable answer has been achieved.

A certain decrease in the quantity of transmitted features occurs due to the masking phenomenon. Since the observed processes, from the isolation of features of signals to the making of decisions, are accomplished by means of essentially non-linear transformations of the values characterizing the stimuli, such events take place as the suppression of a weak signal by a strong one when the two act simultaneously, etc.

When we study the spectra of speech sounds (especially those isolated from running speech), we easily find that they contain a large quantity of maxima, minima, sharp decreases, etc. As is demonstrated by psychoacoustic experiments, not all of the isolated maxima have always the same significance for classification. Furthermore, in the process of being isolated mnay of them are suppressed by more powerful neighbors. Up to now there is insufficient information for establishing the rules according to which certain and not other features (e.g. the spatial distribution of the remaining maxima) are selected by the nervous system in the solution of a concrete classification task. However, the possibility itself for the existence of an operation of surveying various combinations of features appears very probable.

## 4. The auditory description of the speech signal

The question about the form of the auditory description of the speech signal, i.e., about the set of features and their nature, on the basis of which the signal is characterized in the process of perception, has been left practically unexplored up to now by psychoacousticians and linguists. Almost all information concerning the acoustic properties of speech sounds and their perception has been obtained either by the analysis of differences in dynamic spectrograms of natural speech signals having a different phonemic value, or by means of investigations of the perception of synthetic speech-like stimuli. The method according to which an investigator describes the signals is in both instances predetermined by the properties of the instruments with which he is working. As a result, the terminology used for the description of the features of the speech signals has turned out very specialized and at the same time poorly formalized. This refers to such basic concepts as formant, transition, locus, burst, etc.

The qualitative peculiarities, on the basis of which one may distinguish different speech sounds from each other on spectrograms, are more or less known at the present time. Thus, in order to determine which consonant starts a simple CV-syllable, it is useful to examine the following features of the signal [31, 32, 33, 71]:

1. Presence or absence of the fundamental frequency from the very beginning of the syllable;

2. Abrupt rise in the frequency of the fundamental at the transition from the consonant to the vowel;

3. Presence or absence of noise from the very beginning of the syllable;

4. The frequency position of the spectral maximum of the noise segment or the burst;
5. The bandwidth of the noise:
6. The slope of the increase in intensity of the noise:
7. The duration of the noise segment:
8. The intensity of the noise;
9. The presence of formant structure from the very beginning of the syllable;
10. The lack of discontinuities in formant structure during the extent of the syllable.

Besides that, the so-called formant transitions are of essential value in determining the point of articulation of the consonant and for determining whether it is 'soft' or 'hard' (in Russian).

Even though the listed features have not been formalized (it is not indicated by which methods the features may be detected, nor have decision-making rules been given), extablishing the list itself constitutes an important stage in the investigation of speech--the stage of acquiring primary acquaintance with the acoustic peculiarities of the speech signal.

If such a qualitative and, consequently, not very definite description of the speech signal is satisfactory for a number of phonetic tasks and is fairly popular among phoneticians, engineers prefer to use a completely formal, but very inconvenient 'complete' description of the signal in frequency and time. In this process the dimensions of the space in which phonemic decisions are made turn out to be very large, and the decision functions are complex. In addition, such basic difficulties appear as the inevitability of a very precise preliminary segmentation of the speech stream and the indispensable necessity of time normalization.

The hypothesis in favor of which data will be presented below consists of the following. We propose that the complete description of the signal takes place in the cochlea alone; in the processing and transmission of the signal in the neural net, a gradual reduction of information takes place. At a certain level, a transition takes place from the description of the whole envelope curve of the spectrum at a given moment in time (the curve describing the distribution of impulsation along the projection of the cochlea) to the description of the position on the frequency axis (the projection of the cochlea) of a few points (maxima or turning-points) on this curve. On a still higher level, a determination of the parameters of the curves takes place, which describe the displacement in time of the points isolated along the frequency axis. Such assumed parameters might be the direction of movement, speed, magnitude at the turning-point, etc. It is assumed that these already fairly complex characteristics of the signal constitute the features with which the block operates that carries through the phonemic interpretation of the stimulus. It seems to us also that this point of view is not original at all, and that the majority of speech researchers, not discussing it specifically, nevertheless proceed from this standpoint.

For convenience in presenting the material we divide it into three groups of data, concerning the perception of: 1) the envelope curve of a stationary complex signal, 2) the fine temporal structure of a stationary complex signal, 3) changes in time of the spectrum, fine temporal structure, and the intensity of the signal.

## 4.1. The perception of the envelope curve of a complex stationary signal

As a rule, speech researchers use the assumption that in perception, humans somehow determine the formant frequencies of the speech signal and employ this information in making phonemic decisions. Correspondingly, stimuli are described in terms of formants, and the results of experiments are presented from the point of view of the dependence of the response probabilities (identification or differentiation) upon the parameters of the stimulus. Since to the best of our knowledge none of the researchers has observed any contradictions in the obtained results, the description of the stimulus in terms of formant frequencies appears adequate enough.

However, a change in the formant frequencies of a synthetic speech-like stimulus signifies a change in the spectral envelope of this stimulus. Therefore the assumption is not excluded that in fact humans employ in perception not the values of formant frequencies, but either the whole spectral envelope (the outline of the distribution of impulsation along the projection of the cochlea) or, for example, the relative amounts of energy in some fixed bands. The latter point of view was proposed by Varšavskij in the discussion of a possible model of speech perception [34].

In order to prove that a human listener in fact measures formant frequencies, it is first of all indispensable to decide upon a sufficiently formal definition of a formant. Further, it is necessary to find some kind of a reaction which would regularly change when a formant frequency is changed, and would not depend on other parameters of the stimulus.

The term 'formant' is used by speech researchers in two different meanings: a formant is understood to be either a pole in the transfer function of the vocal tract, or a maximum in the spectrum of the analyzed sound. In the latter case, what one really has in mind is a maximum on the curve describing the response of the analyzing apparatus to a given sound. It is understandable that such a definition of a formant is already very indefinite; it seems to depend on the properties of the analyzing device.

If the isolation of formants takes place at a relatively low level in the processing of auditory information, it is difficult to assume that complex procedures are employed in the process--procedures making use of data concerning the transfer function of the vocal tract (e.g., procedures of the 'analysis-by-synthesis' type, cf. [35]). Then the first of the two above-quoted definitions of the formant does not apply, and the formant may, as a first approximation, be identified with a spectral maximum or, more truthfully, with a maximum on the curve describing the response to the signal in those initial links of the auditory system which perform the spectral analysis.

In psychoacoustics it is accepted that the response curve is sufficiently well reflected in the curve describing the masking called forth by the stimulus (for discussion, cf. [36]). Questions as to whether the auditory system employs the frequency of a spectral maximum of a complex stimulus as a parameter of that stimulus, and in which way it measures the frequency of the maximum, relate to the general problems of the physiology of hearing and of psychoacoustics and are closely connected with questions concerning the mechanism of auditory determination of frequency (hypotheses concerning these mechanisms are summarized in ref. [36]).

In the works of Šupljakov [37, 38, 39] direct proof was obtained that a human listener determines the value of the frequency of the first spectral maximum in natural and synthetic sibilant (fricative) consonants of the type /s/ and /ʃ/. In the given case this maximum corresponds to the second formant (in the sense of pole). It appeared that the frequency of the maximum carries two kinds of information: on the one hand, it determines the musical pitch of the sound and on the other hand, the phonetic category ('hard' or 'soft' consonant) to which a given sound belongs. The connection between the frequency of the maximum and the pitch of the sound may be considered immediate; the decision regarding the 'hardness' or 'softness' of the sound is determined by whether the frequency of the spectral maximum is higher or lower than a fixed threshold. The following constitutes proof that in this case it is the frequency of the spectral maximum that is determined and not some other parameter of the spectral function: a change in the amplitude of the maximum has no significance as long as it is above the detection threshold, and the value of the boundary between 'hard' and 'soft' consonants on the basis of the maximum is the same for /s/ and /ʃ/, which are significantly different from each other on the basis of other peculiarities of their spectra. The precision of the determination of the position of the boundary by subjects employing the method of active search [39] appears very high (1.5 - 3.0%), which suggests that the procedure for auditory isolation of the maximum and for determining its frequency is sufficiently effective.

Data obtained in the investigation of the perception of synthetic whispered vowels [40] indicate that the frequency of a spectral maximum corresponding to the first formant of a vowel is also determined. It was discovered that the boundary between the vowels [i] - [o] and [ʉ] - [ø] in the $F_1$-$F_2$ plane is represented by a straight line, parallel to the axis of the second formant. This means that for separating vowels according to these categories, what is employed is the frequency of the spectral maximum corresponding to the first formant, and not the whole spectrum envelope or, for example, the ratio of energy in some fixed frequency bands.

The data quoted above were obtained for the case in which the concept of spectral maximum is not in doubt (the maximum is sufficiently sharp), and the frequency of the spectral maximum coincides with formant frequency defined as a pole in the transfer function of the vocal tract. The question naturally arises what is being employed as parameters of the spectral function, if the range of isolated frequencies (i.e. frequencies containing the essential part of the energy) is sufficiently wide and does not possess a clearly defined maximum. Some data for replying to this question have been obtained in experiments dealing with the estimation of the pitch of band-limited noise. It has been found that for noise whose spectrum is limited only from one end (high-frequency or low-frequency noise), the perceived pitch is determined by the frequency of the cutoff [38, 41]. For band-passed and relatively narrow bands of noise, pitch is determined by the average geometric frequency of the noise [42, 43]. It is obvious that a further special investigation is needed concerning the width of the band at which the transition from one to the other mode of pitch estimation takes place.

It is a very complex and as yet unexplored question as to which parameters are used for describing the response of the auditory spectral analyzer to a signal with a discrete spectrum (stationary vowels). On the basis of available psychoacoustic data [44] one must expect that when the fundamental frequency of the voice is very low a separate maximum on the response curve of the analyzer must appear corresponding to almost each harmonic in the frequency space below 1000 Hz. There is not one of these maxima that may not coincide with a formant frequency in the sense of a pole in the transfer function. Ths supposition that a human listener perceives the frequency of the strongest harmonic as the formant frequency does not agree with data about the great precision in distinguishing the frequency of a formant (as a pole). According to Flanagan's data [45], the just noticeable difference in the frequency of the first formant amounts to 3%. Thus this question remains unclear at the moment and urgently demands further investigation.

## 4.2. The perception of the fine temporal structure of a stationary complex signal

In the perception of speech, the decision regarding the character of the source of excitation (voiced, noise-like, mixed, impulse-like) is made on the basis of the fine temporal structure of the signal. Discrimination between consonants which differ among themselves on the basis of the source of excitation is practically not disturbed at all under significant spectral distortions of the speech signal [46, 47]. The question is still open as to which parameters are employed in the auditory system to describe the temporal structure of sound. It is assumed that they must be some parameters of the distribution of intervals among nerve impulses.

When speech sounds are to be classified according to their temporal structure, it is convenient to divide them first of all into two groups on the basis of whether their structure can or cannot be auditorily determined. Existing data allow one to believe that if the duration of the first of a sequence of two adjacent stimuli is shorter than 15-20 msec, a listener does not recognize its temporal structure. When either fricatives (like s) or periodic consonants (like m) are shortened up to that duration, a listener perceives both as plosives (p or t), which are characterized in speech production by impulse-like excitation [48, 49].

Within the class of stimuli with a temporal structure that can be auditorily distinguished, three categories may be established: periodic signals (with a harmonic spectrum), amplitude-modulated noises, and continuous noises. The amplitude modulation of noise in voiced fricative consonants approaches right-angled modulation (for a discussion of noise production mechanisms, cf. [50]); its frequency is fairly low and corresponds to the frequency of vocal-fold vibration. Psychoacoustic experiments (cf. the surveys in [36, 51]) demonstrate that at these modulation frequencies listeners not only detect them, but distinguish one modulation frequency from another and, furthermore, assign a pitch to the signal that is equivalent to the frequency of modulation. It has been also shown by numerous experiments (cf. the survey in [36]) that the period of repetition of a complex periodic vibration is perceived and serves as the basis for estimating the frequency of the sound, even if the spectrum of the sound does not contain corresponding low-frequency components.

One may thus assume that for the description of the fine temporal structure two parameters are necessary and probably sufficient. One of them must reflect some kind of measure of the degree (explicitness) of periodicity, the other must reflect the magnitude of the period of repetition.

In psychoacoustics, perception of the temporal structure of the signal is sometimes taken to include the perception of changes in the signal that occur with a frequency below that of the vibration of the vocal folds (below 100 Hz). Periodic low-frequency changes in the signal are almost never encountered in speech (the exception is provided by the sound r); however, single occurrences of rapid changes in the spectrum (formant frequencies), fundamental frequency, or intensity level appear regularly. It seems to us that the parameters by means of which these changes are described can be considered as derived from the parameters that have been examined in this section (formant frequencies, period, degree of periodicity). The next section will be devoted to their consideration. The intensity of vocal fold vibration should obviously also be assigned to parameters of the same level as formant frequencies and the period of repetition. The time constant for the auditory determination of intensity is approximately 10-20 msec [52, 53].

The essential characteristic of the parameters listed above is that auditory measuring devices responsible for their detection must have a low inertia. Therefore it is possible to consider these parameters as indicators of instantaneous (current) properties of the speech signal. Only stationary sounds (synthetic or artificially pronounced isolated vowels and some consonants) can be described with a single value for each parameter. In natural connected speech, the values of the parameters of the signal constitute functions that change with respect to time.

### 4.3. The perception of changes in time of the spectrum, fine temporal structure and intensity of the signal.

An essential characteristic of natural connected speech is the fact that the values of formant frequencies, fundamental frequency and intensity change significantly in the course of the duration of separate speech elements, and that these changes are by no means random, but have a completely regular character. From the point of view of speech production, the stream of speech may be considered as a sequence of open syllables [14, 54, 55, 56]. A standard kind of time function corresponds to each syllable and each parameter (formant frequencies etc.), as well as a limited set of possible transformations of the function, connected with the tempo of pronunciation, intonation, and position of the syllable within a word.

This means that the curve describing the change of each of the parameters in the course of a long utterance may, as a first approximation, be viewed as a sequence of sections (pieces), where each section is a certain standard time function corresponding to a syllable.

The question is almost unexplored as to how a signal is described in the process of perception whose parameters change in time. In order to discuss the small amount of fragmentary data that are available it would be useful first to summarize existing hypotheses.

One of the hypotheses says that the time picture is described completely, i.e. that readings are used for each of the parameters taken at, for example, every 10 msec. Thus the change in the parameter during the extent of the syllable is described by a set of numbers reflecting the value of the parameter at successive discrete instances in time. The difficulties connected with this hypothesis consist first of the fact that such a description appears extremely unwieldy (a large memory capacity is required for registering it) and, secondly, of the fact that phonetic and prosodic information has not yet been separated.

When this form of description is employed, the same syllable produced by the same speaker will look different, depending on the tempo with which it was pronounced, on the position of the syllable within a word, or whether it carries logical stress, etc. Therefore it is still difficult to use such a description as the immediate input to the block that performs the phonemic interpretation.

Two ways have been proposed to overcome these difficulties. In one of these [57], the representation of the syllable obtained at perception that has been entered in short-term memory is subjected to certain (for example topological) transformations, as a result of which it is given a more standardized shape. The formalized representation enters at the input of the block performing the recognition.

The second way [58] consists of comparing the representation entered in short-term memory with a standard syllable which is synthesized under the assumption of a different tempo, position within the word, etc. The phonemic composition of the syllable, tempo, position within the word etc., constitute initial variables for the synthesizer. What is being sought is such values of those variables with which the synthesized representation and the representation entered into short-term memory are closest to each other.

Although the presented hypothesis of recognition by syllables appears sufficiently logical from the point of view of existing knowledge about the process of speech production, there is no proof whatsoever that perception of speech by humans is indeed accomplished in this manner. Furthermore, there exist data that contradict this hypothesis. These data concern the possibility of partial recognition of the syllable, recognition of some of its phonemic (distinctive) features while others are not recognized (for example, the recognition of the manner of articulation of a consonant without recognizing its point of articulation, etc.), and recognition of prosodic features without the recognition of phonemic ones.

Into this category falls also the fact that some phonemic features can be recognized earlier than others, even before the listener hears the complete syllable [14]. These data suggest that the auditory description of the syllable that enters at the input of the block performing the phonemic interpretation is already organized in such a manner that it allows parallel, multi-channel processing.

The second hypothesis consists of the proposition that the curves which reflect the change of formants, fundamental frequency, etc., during the extent of the syllable are described in perception by means of the set of features of those curves. These features might include the direction of change of the parameter, the rapidity of change, and the value of the parameter at a certain specific point.

In experiments in the perception of synthetic stimuli it has been shown that the nature of the initial transitions of the second and third formants of vowels carries information about the point of articulation of the consonant [60, 61, 62, 63]. At first it was proposed that the characteristic of the transition used by human listeners is its 'locus'--the initial value of the formant frequency, which supposedly does not depend on the vowel with which the given consonant is connected in the CV syllable. (Later it was discovered that the locus value is different for different vowels [64]).

Not long ago Stevens [65] examined in detail spectral changes in the sound during the transition from stop consonants to vowels at different points of articulation of the consonant. Comparing his results with data concerning the perception of formant transitions obtained at Haskins Laboratories, he advanced the hypothesis that when formants are close to each other in frequency, what is determined in perception is the frequency position of the sum of the spectral maxima. From this point of view, the transition from labial consonants to vowels is always characterized by a rise in the frequency of the spectral maximum, and the transition from apical consonants to vowels is characterized by a lowering of the frequency of the maximum. As the absolute frequency position of the maxima depends on the nature of the vowels, it is natural to admit that the useful feature which distinguishes between labial and apical consonants is the direction of the change in time of the frequency of the maximum.

Data in direct support of the assumption that human listeners use the direction of change in the frequency dimension of the energy maximum as the indicator of the point of articulation of the consonant were obtained in experiments for determining the boundary between *p* and *t* in synthetic syllables constructed in the following manner: the syllable consisted of a short (10 or 15 msec) emission of noise with a narrow bandwidth or of a sinusoid, followed by a harmonic signal with a maximum in the range of 300-600 Hz (*u*) or 900-1200 Hz (*a*). The subject changed the frequency of the tone (the center frequency of the band-passed noise) in the short emission, trying to find the value at which perception changed from *tu* to *pu* (*ta* to *pa*) or from *pu* to *tu* (*pa* to *ta*). It turned out that the boundary between *pu* and *tu* is located near 400 Hz, and the boundary between *pa* and *ta* around 1000 Hz [66].

The notion that it is the direction of the change that is perceived and not the locus, i.e., the initial value of the formant frequency, is supported by the categorical nature of perception [67]. Subjects behave as if they detected only the presence or absence of a change in the formant and its positive or negative sign.

The transition from a velar consonant (*k*, *g*) to a vowel is characterized by the fact that the frequency of the maximum does not change its position in time, but the pertinent frequency range at the beginning of the transition is narrow (the frequencies of $F_2$ and $F_3$ coincide) and then becomes wider (the formants separate during the vocalic segment) [68]. In Russian, one of the features distinguishing *k* - *g* from *t* - *d* and *p* - *b* is the greater duration (for *k* - *g*) of the noise of the explosion.

There exist also data supporting the notion that humans use the rate of change of formant frequency as a useful phonetic feature [69].

If the direction and the rate of change of formant frequency play the role of useful features of consonants, then, it is proposed, useful features of vowels consist of the values of formant frequencies during the stationary part of the vowel (if present) or at the turning-point of the formant curve.

Of great interest are here experiments in estimating the frequency of a short stimulus (20-50 msec) whose frequency was changed significantly (raised or lowered) during its extent [69, 70]. These experiments showed that humans equate the pitch of such signals with the pitch of a stationary tone having a frequency equal or close to the terminal frequency of the signal. These data permit one to exclude the assumption that it is the time average of the frequency of the changing signal that serves as an auditory parameter.

These experiments should be followed by an investigation of the perception of the pitch of sounds with more complex changes of frequency in time.

Fairly little is known regarding the auditory features employed in describing the curve that represents the change of the fundamental frequency of the voice during the extent of the syllable.

It has been shown that a sudden jump in the frequency of the fundamental at the transition from consonant to vowel serves as a useful feature in distinguishing *b* from *m* [71]. The threshold value for the rise in fundamental frequency that corresponds to the boundary between these consonants consists of 10% of the absolute value of the fundamental frequency.

On the basis of a fair amount of phonetic data it may be assumed that a more gradual rise (proceeding with lesser rapidity) in the fundamental frequency during the syllable serves as a feature of stress.

Data are available showing that the relative intensity of a consonant constitutes a useful feature [72] for distinguishing fricatives from each other [71]. However, the question has not yet been investigated at all which and how many parameters are used to describe auditorily a signal changing with respect to intensity.

In conclusion it is indispensable to turn to one more parameter that obviously must be assigned to the same level in the processing of the signal. This parameter is the duration of the section (segment) of the signal.

The interest in this parameter consists in the fact that it makes obligatory a preliminary segmentation of the utterance into sections.

There is no doubt that a human listener somehow determines the duration of segments corresponding to vowels (it is possible to determine stress placement on the basis of comparing the durations of vowels in a sequence; in a number of languages the duration of vowels has phonemic significance). It is also known that the duration of the hold (approximation) of a consonant makes it possible to distinguish a double consonant from a single one.

There exist experimental data about the discrimination of the duration of the pause corresponding to the closure of a voiceless stop consonant [73, 74], and about imitating the duration of a voiced stop consonant in an isolated CV syllable [49]. On the basis of these data it is possible to assume that the value of the duration of a segment, established by the auditory system, is a monotonic function of its physical duration. In the phonetic interpretation of the obtained value supplementary information is employed, concerning probably the tempo of speaking and/or the duration of other nearby segments.

The question appears completely unexplored regarding the mechanism of auditory segmentation of the utterance. Therefore we can only list some assumptions bearing on this question.

The most obvious of these is the assumption that auditory segments need not coincide with phonemes in the sense that each segment contains information about one and only one phoneme and that the number of segments is equal to the number of phonemes.

One of the possible assumptions is that segmentation is accomplished as a result of processing the complete spectral-temporal description of the signal, and the points of segmentation are established at instances at which significant changes take place in the spectral picture. This would correspond, logically, to assuming that the isolation of segmentation signals proceeds at the same level as the isolation of parameters describing 'instantaneous' values of the spectral envelope and fine temporal structure. The functional meaning of segmentation signals might be that they control the consideration (transmission into short-term memory) of output signals from feature detectors, characterizing the dynamics of 'instantaneous' parameters during the extent of the segment.

Another proposition is that segmentation signals are formed as a result of processing 'instantaneous' parameters, and a separate segmentation might be performed for each of the parameters.

It is probable that the segmentation signals cannot be processed at a rapid rate. Thus, according to psycho-acoustic data, the temporal threshold of the perception of sequential ordering is approximately 20 msec [75].

## 5. The Phonetic Interpretation of Speech Stimuli

### 5.1. Units emerging from the block of phonetic interpretation.

Under the phonetic interpretation of a stimulus we understand the process of working out an auditory description of the stimulus, as a result of which a definite articulatory reaction may be associated with the stimulus. If large numbers of such reactions $R_1$ and $R_2$, observed in response to numerous repetitions of stimuli $X_1$ and $X_2$, do not differ among themselves, we accept that one and the same phonetic description, and one and the same phonetic image corresponds to both stimuli.

The phonetic image may be specified either as the set of instructions for synthesizing the speech complex in case we consider it from the point of view of the final stages of transformation in imitation, or as the designation of a multitude of stimuli (and a multitude of auditory descriptions) possessing certain given properties, in case we consider it from the point of view of initial stages of transformation [75].

Inasmuch as the phonetic image is an abstract description of both the acoustic stimulus and the motor complex, its internal structure must reflect the constraints that are essential both to the auditory system and to the system of speech production.

At the basis of contemporary linguistic investigations of language lies the assumption that the speech signal is described in perception and production in terms of a set of segmental units--phonemes, and suprasegmental units--prosodemes. This assumption is supported by a series of experimental data. Thus, a study of the mimicry of vowels [76] showed that in response to a signal, the subject selects one of a limited set of known configurations of the vocal tract. Thereby a certain category (multitude) of speech stimuli corresponds to each configuration, so that information about the required configuration may be represented in the phonetic image in the form of a symbol.

The contradiction, well known to engineers, between the linguistic approach to a phonemic system and the 'technical' (from the point of view of automatic speech recognition) description of phonetic images consists in the fact that the set of phonemes must be minimally small for a linguist, while the set of phonetic images need not meet this condition for an engineer. The requirement for economy may be left unsatisfied if it counteracts the requirement for reliability in recognition.

Cases in which the sets of phonemes and phonetic images do not coincide are found in instances in which one and the same phoneme is realized in essentially different ways as a result of the influence of immediate phonetic context. As the most characteristic example of this we may consider Russian vowels after hard and soft consonants: ta - t'a, to - t'o, etc. [77, 78].

From the point of view of the reliability of automatic recognition, it is useful to describe separately the groups of vowels after hard and soft consonants and to assign to them separate symbols. An experimental study of the perception of these vowels showed that listeners (native speakers of Russian) proceed in exactly this manner--they interpret the a of ta and the a of t'a as separate entities, although from the linguistic point of view they constitute one phoneme /a/ [78].

At the present time enough data have been accumulated [83, 84] to maintain that the number of different entities used by the brain of a native speaker of Russian in the interpretation of vowels is larger than the number of vowel phonemes in the Russian language established at the linguistic level.

In order to designate these entities one might want to introduce some kind of new terminology, since they do not coincide completely with phonemes determined linguistically. The essential difference between them is that basically, from a linguistic point of view, redundant features of phonemes (those that arise, for example, as a result of the influence of some other phoneme) are not considered important for their description and isolation as separate phonemes, whereas for a listener it is indeed the redundant features that are made use of. In the following discussion the term 'phoneme' will be used, but the term will be understood to refer to the subjective image employed by the brain of the listener in the process of speech recognition. Other investigators use the term 'sound type' in this sense.

It has been shown experimentally that a number of subjective images--phonemes--really exists in the human nervous system, and that this number is not only finite but quite limited in size. Final data about the size of this number for native speakers of each concrete language are not yet available. As was mentioned above, only the minimal set of phonemes is established linguistically; it is unclear, however, whether any arbitrary linguistic phoneme can be associated with a phonetic image. For solving this problem it is indispensable to turn to methods of experimental psychology that have been worked out in the last few years (the method of mimicry, the analysis of confusion matrices, the method of active search for boundaries along phonetic categories, and methods of psychological scaling).

Let us consider how one might describe a phoneme taking it as a symbol denoting a certain range of auditory images and a certain articulatory complex. According to one of the methods for producing such a description, each one of the symbols is independent of every other symbol. According to another approach, the set of phonemes is systematically arranged, and each phoneme is described by listing the values of some of its 'distinctive' features. In this case, the number of features is significantly smaller than the number of phonemes. The idea of such a description of phonemes belongs to N. S. Trubetzkoy

[84]; it has been developed by R. Jakobson, M. Halle and G. Fant [31].

The logic of such propositions is rather obvious, if one looks at the phoneme as a set of instructions for the synthesis of an articulatory complex. These instructions must relate to speech organs (groups of muscles), and they can be considered as a set of elementary commands addressed to the various organs (vocal folds, lips, different muscle groups of the tongue etc.). One of the basic tasks of present-day physiological phonetics is to determine which sets of elementary instructions (motor commands) correspond to phonemes [85].

The idea, formulated by linguists, that phoneme sets are inherently systematic, also finds confirmation in specifically linguistic regularities (positional and conditioned sound changes, historical sequences of changes, morphological regularities, etc.). The description of such regularities appears more economical, if a phoneme is represented not as an isolated symbol, but as a list of its distinctive feature values.

Experimental proof that human listeners recall phonemes on the basis of a set of feature values was produced in investigations by Wickelgren [86, 87] and Galunov [82]. In these experiments listeners were presented series of six sound sequences, e.g. of the type CVC, where the vowels were different, but the consonants remained the same. The subjects had to write down the sequences after having heard the whole series. Mistakes made in writing down the recalled sequences were analyzed. It turned out that the mistakes have a very regular character. For each transmitted phoneme there exist some 'close' phonemes with which it is most frequently confused. This could not be the case, if the phonemes were remembered as isolated symbols, unconnected with any other symbols--phonemes.

Thus it is advantageous to accept that phonemic information, as it emerges from the block of phonetic interpretation, must be represented in terms of abstract distinctive features. Which must be the concrete set of these features and how many gradations are possible for each feature remains as yet unclear.

Very important is the question concerning the mutual connections between acoustic (auditory) features of the speech signal and the distinctive features of phonemes. The simplest and most attractive, although as yet experimentally unproven, is the proposition that distinctive features are binary, and that for each distinctive feature there exists a corresponding decision rule, its proper decisive boundary in the space of auditory features. If the auditory image that is called forth by the stimulus is located to one side of the boundary, the value of the stimulus according to the distinctive feature has one sign, and if it occurs on the other side of the boundary, its value according to the distinctive feature has the opposite sign.

It has been established sufficiently firmly at the present time that information concerning one and the same distinctive feature is contained in several acoustic (auditory) features of the stimulus [63, 89, 9]. This means that the decisive boundary may constitute a hypersurface in the space of these several auditory features. If the auditory features themselves are binary (cf. the preceding section), the decisive boundary may have the maximally simple form

$$\sum_{1}^{i=n} k_i x_1 = 0,$$

where $x_i$ - the auditory feature (+I or -I), $k_i$ - the weight of the given auditory feature.

Recently obtained data concerning the decision rules that are employed in distinguishing between synthetic b and m [75] fit into this kind of a primitive scheme. It was also discovered that a human subject is able to give a numerical estimate of the closeness of the synthetic stimulus to the phoneme [75]. This makes it possible to admit that information concerning the distinctive feature at the output of the phonetic interpretation block determines not only the sign of the function

$$\sum_{1}^{i=n} k_i x_1,$$

but also its modulus. This is equivalent to saying that what is remembered is not a categorical decision concerning the class of phonemes (e.g. nasal or non-nasal) to which a given stimulus must belong, but the probability with which the stimulus may belong to this class.

The advantages of preserving this kind of information have already been discussed above (section 2.2.). It was experimentally proven by Lindner [90] that a final phonemic decision concerning an uncertain vowel may be made after the second vowel following the first one in time has been perceived.

## 5.2. The procedure of phonetic interpretation of auditory descriptions.

Most complex appears to be the question concerning the temporal organization of the process of phonetic interpretation. Direct experimental data concerning this question do not yet exist; however, some important requirements are known which must be met. One of them is that the procedure must ensure the collection of information that is contained in auditory features of different nature, distributed in time within the limits of approximately one syllable (cf. the surveys in [14, 90]). A second and most important requirement is that the failure to recognize an element must not lead to its being omitted--it must be indicated in the completed sequence of phonetic images that at a given point within the sequence there was an unrecognized (partially recognized) element [90].

The first of these requirements presupposes the existence of memory. Information regarding distinctive features (let this be the meaning of the function $\sum k_i x_i$) must accumulate with each occurrence (in the general case, non-simultaneous occurrence) of the auditory features $x_1$, $x_2$, ... $x_n$. The second requirement presupposes an obligatory segmentation and breaking-up of the data. In the contrary case, information about the first, not yet completely recognized element will be mixed with information about the second element in the temporal sequence. As a result the first element will be left out, and the second may be incorrectly determined. It seems to us that one of the most important tasks for the immediate future consists of working out several different models for procedures that would satisfy both above-mentioned requirements, and of finding methods for their experimental verification.

Below we will attempt to describe--as yet in a very tentative manner--a hypothesis that appears plausible for a series of reasons, and, according to our opinion, requires further elaboration and testing. It might be called the hypothesis of syllable recognition.

We propose that the process of phonetic interpretation includes the operation of a special program that marks off syllables (open syllables). It is clear that the isolation of elements constituting rhythmic and melodic structures takes place during the perception of very varied and not even necessarily speech-like signals. The rhythmic and melodic structure of a phrase may be transmitted by means of signals that are very remote from speech sounds. There exist data that some patients with sensory aphasia have no difficulty in reproducing the rhythmic structure of a phrase, although they can neither understand it nor reproduce the sequence of phonetic elements of which it is constituted [6].

Under very high spectral distortion and correspondingly very low phonemic intelligibility of speech, the perception of its rhythmic structure is almost unaffected [90]. This makes it very plausible to assume that in speech perception two independent procedures are employed in parallel. One is responsible for the segmentation of the stream of speech into syllables (elements of rhythmic sequence) and the description of the so-called prosodic characteristics of the sequence, the other is responsible for the description of the characteristics of each separate syllable, which is accomplished in terms of phonemes or distinctive features.

We just used the term 'segmentation' of the stream of speech. Since it is frequently used in very different meanings, it is indispensable to dwell somewhat more specifically on what we have in mind. As of now we propose only that as a result of some kind of a procedure every syllable is associated with a kind of 'mark' (impulse), so that the number of impulses that arise in the process of listening to a sentence will be equal to the number of syllables in that sentence.

From the fact that a human is able to repeat a meaningless sequence of 7 - 10 syllables without confusing their order in time and without distorting the prosody, it follows that when perceived information is registered in memory certain reference signals must be employed that allow one to group together phonetic and prosodic information about the syllable and assign the syllable its order number. The role of such reference signals must be played by the proposed syllable impulses. It is possible, for example, to assume that the short-term memory into which the speech sequence is entered consists of K cells which are filled in sequence. The syllable impulse performs the role of a switching signal, switching the output of the preceding level of the system from one cell to the next.

Allowing for obligatory switching makes it possible not to make final phonemic decisions if sufficient information is not available during the extent of the syllable; it makes the possibility of prolonged preservation of information regarding the input stimulus compatible with the absence of confusion with data relating to successive phonemes.

Up to now we have said that a separate memory cell corresponds to each successive syllable. However, we have also said that information within the cell is entered in terms of phonemes. Output signals of the preceding level correspond to a running description of the stimulus in terms of auditory features. In order to proceed from one kind of description to the other it is necessary to use some kind of a decoder.

Our next proposition is that within the nervous system there exists a series of identically organized decoders (their number is equal to the number of syllables that can be kept in memory simultaneously). The syllable impulse accomplishes the successive switching of input information from one decoder to the other. Each separate decoder accomplishes the transition from the sequence of auditory features present during an open syllable to the description of this open syllable in terms of phonemes or distinctive features.

The fact that the decoder must be designed to operate on open syllables follows, firstly, from the fact that speech is articulated as a sequence of open syllables (the articulation of the vowels begins simultaneously with the articulation of the consonant [91, 92]) and, secondly, from the observation that the interpretation of the stimulus during the consonant part depends on the properties of the stimulus during the part of the following rather than the preceding vowel [14, 93].

The use of open syllables in the capacity of input signals to the decoder appears very reasonable both from the point of view of the procedure of phonemic recognition and from the point of view of the relatively low requirements to be made in this case with regard to the procedure of segmentation. The collection of information about the phoneme may be performed during the whole temporal segment in which this information is actually present; stationary and transitional parts may be utilized equally. The number of elements in the output alphabet of the decoder may be approximately equal to the number of phonemes, since the contextual mutual influences may be accounted for in the decision rules themselves.

According to the motor theory of perception, the work of the decoder consists in transforming the perceived signal into a set of motor commands which would be required for imitating what is heard. How is the selection of required motor commands carried through? It is difficult to suppose that this is done by the method of surveying the complete set of hundreds of thousands of possible variants of syllables. A parallel survey would demand a great expenditure of functional elements, and a sequential one--a great amount of time.

It is possible to decrease the number of variants on the basis of a preliminary recognition of units which are simply connected with elements of motor complexes. These units, obviously, must be close to phonemes. It may turn out that the reliability of recognizing phonemes on the basis of their acoustic characteristics will not be high. It is possible to imagine that in this case certain of the most probable phonemes will stand out with the indication of their probabilities. After this, motor complexes are formed which are required for imitating sequences of these most probable phonemes. The number of possible variants of such sequences will be smaller by several orders of magnitude than the initial number of possible variants. From these variants the variant will be selected that possesses the maximal production values for three quantities: the a posteriori probability of phonemes, their a priori probability, and the a priori probability of the sound sequence. The last two values reflect knowledge of the laws of the language. In this manner decisions about the recognition of sound types are made more precise simultaneously with the recoding of the acoustic signal into motor commands, i.e., into a very compact code.

The procedure described here may explain one of the peculiarities of motor perception of speech. At the same time it is a short description of the above-mentioned algorithm for increasing the reliability of recognition on the basis of the redundancy of the signal [21].

Let us return to segmentation. The basic requirement to be set up for this procedure is that information pertaining to a phoneme in one syllable must not be attributed to a phoneme of another syllable. It is granted that the syllable impulse arises somewhere in the transition from vowel to consonant. Omission of the transitional section is not dangerous from the point of view of recognizing the consonants of the second syllable, since the transition contains very little information [79, 83]; the possibility of attributing this section to the consonant of the preceding syllable can be easily excluded on the basis of limitations incorporated in the schema of the decoder itself (the consonant following the vowel is excluded).

It seems a priori obvious that the most complicated task in working out a model of this type is the recognition of consonant clusters. From this point of view it appears extremely important to obtain experimental data regarding the perception of consonant clusters in nonsense sound sequences.

## 6. Scheme of a Model of Speech Perception

Basing ourselves on all facts and assumptions discussed above, let us now attempt to outline the most plausible general scheme of a mechanism intended for the recognition of a sufficiently large quantity (2 - 3 thousand) of spoken words. A speech signal, constituting a non-stationary function of time $f(t)$, arrives at the input of the mechanism. The output block must issue a decision about the assignment of the unknown realization to one of the earlier indicated 2 - 3 thousand words of the lexicon $S_o$ with a reliability $P_o$, which is comparable to the reliability with which these speech signals are perceived by human listeners. It is clear that this mechanism must have a hierarchical structure of the type represented on Fig. 1. It is indispensable to make more precise the number of stages (elementary automata), make more concrete the contents of each stage and describe the procedure for processing the signal in its progress from the input to the output of the mechanism.

It follows from the foregoing that at each hierarchical level a block may be isolated that carries through the perception procedure (receptor $X_i$), a decision-making block (classifier $D_i$) and a block stating the decisions made with the reliability $P_i$ (effector $S_i$).

It may be expected that because of the limited abilities of the classifiers $D_i$ the recognition of elements $S_i$ will be performed with an unacceptably low reliability $P_i$. It would be useful to have at every level blocks ($H_i$) for the correction of errors. Errors may be eliminated on the basis of a priori information about speech and language, which may be stored in long-term memory. This information is of different kinds--it may constitute knowledge about limitations in the physical characteristics of the speech production apparatus or about linguistic regularities in the language.

Taking into account what has been said, the procedure for recognizing elements at one of the levels may look like the following [95].

The classifier $D_i$ indicates some hypotheses $S_i$ to which the vector of unknown realization $X_i$ may be attributed with the greatest probability. It is logically inevitable that a certain block $Q_i$ (let us call it 'supervisor') be present, which must evaluate the quality of decisions being made and, according to necessity, include reserves of one kind or another for increasing the reliability of recognition. The evaluation of the quality may consist in the simplest case of the determination of the difference $\Delta P_i$ of the a posteriori probabilities of competing hypotheses $S_i$. A decision is considered satisfactory, if $\Delta P_i$ exceeds a certain fixed threshold $V$.

Increasing the reliability may be achieved by providing a more and more complete description of the realization $X_i$, i.e., by the analysis of a wider range of parameters. After that, when these possibilities have been exhausted, and $\Delta P_i < V$, the supervisor includes block $H_i$ (the error-correction block) on the basis of a priori information about the characteristics of the speech tract or linguistic regularities.

As Wald has shown [96], this sequential procedure for increasing the reliability of recognition ensures minimal mathematical expectation of cost for making the decision.

It should be mentioned that the order in which these or other means for increasing the reliability of recognition are adopted depends on the relationship between the useful result of a given method and the cost of its realization. This relationship is as yet unknown to us; therefore the set and the order of inclusion of means by the supervisor may differ from what has been described above.

If at any step the probability of a hypothesis exceeds the probability of any other hypothesis by more than V, then this hypothesis $S_i$ is transmitted to the input of the following (i + 1th) stage of recognition. In the contrary case several of the most probable hypotheses are retained in memory, which are then transmitted to the input of the i + 1th level one after the other, in the order of decreasing probability. It is possible to imagine another--parallel--method, according to which all competing hypotheses enter simultaneously at the inputs of several classifiers of the same type at the i + 1th level.

If in the sequential scheme economy is achieved with respect to the number of functional elements, then in the parallel scheme the time needed for decision-making is decreased. The effectiveness of parallel application of algorithms for solving complex tasks in computing systems [97] points to the usefulness of a parallel scheme for processing information, especially when it is necessary to obtain high productivity on the basis of slowly acting functional elements; however, we do not yet know direct experimental facts in favor of one or the other scheme for the processing of information by the brain.

Facts presented in the beginning of this work speak in favor of the assumption that on the level at which phonemes are recognized, a decision is made taking into account information scattered over a segment of the type of an open syllable. It follows from here that in the scheme of the automaton there must be a block for the segmentation ($C_i$) of the stream of speech into open syllables. It is probable that blocks devoted to the segmentation of the speech stream into one or another kind of sections must be present also at other hierarchical levels. Thus besides X, D, S, P there must be present blocks Q, H, and C. What will the procedure of processing the speech signal now look like, when it passes through the recognition mechanism? In the sequential variant (Fig. 3) the speech signal f(t) is transformed at the very beginning into a rather complete description in the space (X) of frequency and time. During a section ('window') of a certain duration, determined by the short-term memory capacity of the input chains of the auditory analyzer, some features ($S_1$) are isolated of the type of static and dynamic characteristics of formants, characteristics of the noise part of the spectrum etc., normalized for loudness, tempo and some other parameters. It is possible that this procedure is articulated into a series of smaller stages, as, for example, loudness normalization, isolation of static characteristics, tempo normalization, determination of dynamic characteristics, etc.

In technical models the segmentator $C_1$ may be needed for establishing the boundaries of the temporal 'window'.

The indispensable reliability of recognition ($P_1$) of features may be obtained by using information about physical laws of speech production of the following type: the frequency of the fundamental cannot be changed faster than at a certain speed; simultaneous existence of such and such features is impossible: after a given

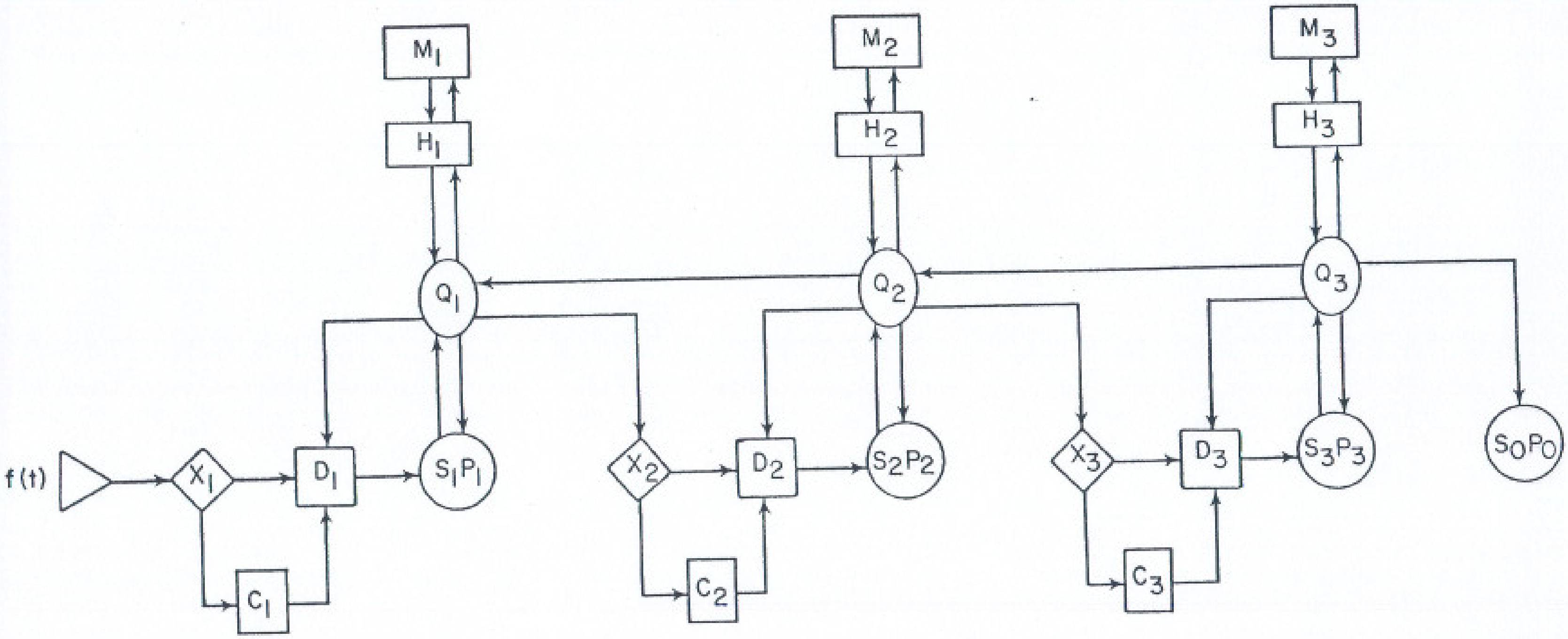

f(t)
$X_1$
$D_1$
$C_1$
$S_1P_1$
$Q_1$
$H_1$
$M_1$
$X_2$
$D_2$
$C_2$
$S_2P_2$
$Q_2$
$H_2$
$M_2$
$X_3$
$D_3$
$C_3$
$S_3P_3$
$Q_3$
$H_3$
$M_3$
$S_0P_0$

Fig. 3

combination of features, the occurrence of another given set of features is most probable, etc.

In living systems this information is probably included in the construction of blocks that measure the characteristics of a speech signal in the form of time constants, bandwidths, schemes for suppressing or sharpening various maxima etc. In technical mechanisms, information about speech production may be stored in the long-term memory $M_1$.

The use of this kind of information is continued until the probability of a certain variant of features becomes greater than the probability of other variants relative to a certain threshold value $\nabla$. This variant enters at the input of the next stage of transformation, where the sequence of such features constitutes the space of description $X_2$.

If the difference between probabilities is smaller than $\nabla$, then several variants of features S1 are kept in memory.

On the second level the recognition of phonemes ($S_2$) takes place. For this purpose, the classifier D2 employs information from the open syllable type segment, whose boundaries are determined by the segmentator C2. For reducing the number of possible variants of a certain phoneme, information is used that is contained in the description X2, and afterwards, if necessary, also information from $M_2$ concerning the structure of phoneme sequences. For this purpose, block $H_2$ formulates sequences of most probable variants of phonemes and, taking into consideration all this a posteriori and a priori knowledge, selects the most probable sequence. If the difference in probability of this selected sequence and any arbitrary sequence exceeds a certain threshold $\nabla$, then the phonemic code of the syllable is transmitted to the input of the following block. In the contrary case, a categorical decision is not made and the phoneme codes $S_2$ of several (most probable) syllables are retained in memory. If there are too many variants, the procedure may be repeated, calling forth another variant of features $X_2$ along the line $Q_2 - Q_1$ at the input of the block.

In order to recognize words from the lexicon $S_o$ the space $X_3$ must contain, in addition to phoneme codes, information about stresses. The segmentator $C_3$ carries through the segmentation of the speech stream into sections stretching from one stress to the next. Two such neighboring sections contain as a minimum one word of the lexicon $S_o$. The search for the needed word and the simultaneous determination of its boundaries may be accomplished by means of the algorithm of Lisenko [24]. At this stage as well as earlier, additional a priori knowledge from $M_3$ about the elements of the lexicon may be used (block $H_3$) in the selection of a decision, and if it should prove indispensable, other variants of the phonemic sequences may be summoned (along the line $Q_3 - Q_2$) to the input ($X_3$).

Differing from this, in the scheme with blocks working in a parallel mode (Fig. 4) several most probable variants of features $S_1$ are simultaneously transmitted to the input of the second level. In each of $\alpha$ branches the classifier $D_2$ (j) establishes whether the vector $X_2(j)$ belongs to one of the phonemes of the alphabet $S_2$.

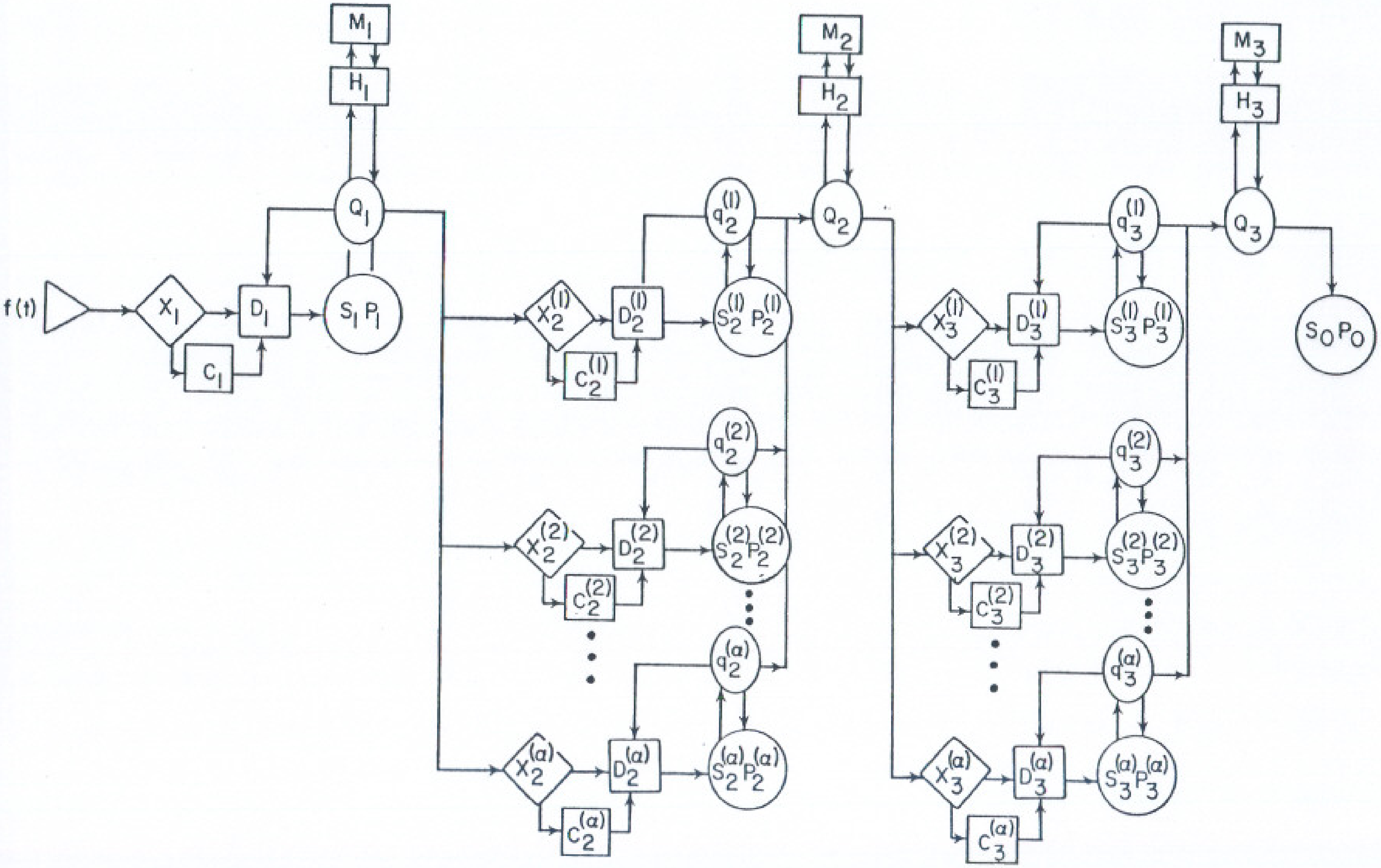

f(t)
$X_1$
$C_1$
$D_1$
$S_1 P_1$
$Q_1$
$H_1$
$M_1$
$X_2^{(1)}$
$C_2^{(1)}$
$D_2^{(1)}$
$S_2^{(1)} P_2^{(1)}$
$q_2^{(1)}$
$X_2^{(2)}$
$C_2^{(2)}$
$D_2^{(2)}$
$S_2^{(2)} P_2^{(2)}$
$q_2^{(2)}$
$X_2^{(\alpha)}$
$C_2^{(\alpha)}$
$D_2^{(\alpha)}$
$S_2^{(\alpha)} P_2^{(\alpha)}$
$q_2^{(\alpha)}$
$Q_2$
$H_2$
$M_2$
$X_3^{(1)}$
$C_3^{(1)}$
$D_3^{(1)}$
$S_3^{(1)} P_3^{(1)}$
$q_3^{(1)}$
$X_3^{(2)}$
$C_3^{(2)}$
$D_3^{(2)}$
$S_3^{(2)} P_3^{(2)}$
$q_3^{(2)}$
$X_3^{(\alpha)}$
$C_3^{(\alpha)}$
$D_3^{(\alpha)}$
$S_3^{(\alpha)} P_3^{(\alpha)}$
$q_3^{(\alpha)}$
$Q_3$
$H_3$
$M_3$
$S_O P_O$

Fig. 4

The most probable hypotheses are transmitted by the supervisors $q_2^{(j)}$ to the input of supervisor $Q_2$, which functions in the same way as $Q_2$ in the sequential variant.

The peculiarities of the functioning of blocks working in parallel at the third level of recognition are analogous.

These schemes do not contradict presently known facts about human speech perception. At the same time we are conscious of the possibility that a further development of investigations in this area may lead either to a concretization of these schemes or to a necessity for changing them in very basic ways.

We believe that the tasks immediately ahead consist of further investigations of the structure, methods of functioning and interactions of human prototypes of the blocks which enter the schemes presented above.

## BIBLIOGRAPHY

1. Gud G.X., Makol R.E. Sistemotexnika. Vvedenie v proektirovanie bol'šix sistem. Perevod s angl. Izd-vo "Sov. radio". M., 1962.

2. Mel'čuk I.A. Avtomatičeskij sintaksićeskij analiz. Novosibirsk, 1964.

3. Lejkina B.M., Nikitina T.N., Otkupśčikova M.I., Fitialov S.Ja., Cejtlin G.S. Sistema avtomatičeskogo perevoda, razrabatyvaemaja v gruppe matematićeskoj lingvistiki. V.C. LGU "NTI", 1966, I, 40-50.

4. Simmons R.F. ECVM, otvečajuščaja na voprosy, zadannye no-anglijski. Zarubežnaja radioelektronika, 1965, 7. 49-82.

5. Galunov V.I., Čistovič L.A.O svjazi motornoj teorii s obščej problemoj raspoznavanija reči. Akustićeskij žurnal, 1965, II, 417-426.

6. Stevens K.N. and House A.S. Speech perception in foundations of modern auditory theories (eds. S. Tobias a.E. Schubert), in press.

7. Fant G. Auditory patterns of speech. Proc. of the Symposium on Models for the Perception of Speech and Visual Form. Boston, Mass., Nov. 11-14, 1964.

8. Chistovich L.A., Golusina A., Lublinskaja V., T. Malinnikova, M. Zhukova. Psychological methods in the speech perception research. Zeitschrift für Phonetik, 1968, in press.

9. Fant G. The nature of distinctive features. Speech Trans. Lab. Quart. Progr. a. Status Rep., 1966, 4, 1-14.

10. Nil'son N. Obučajuščiesja mašiny. Izd-vo "Mir", M., 1967.

11. Miller G.A. The magical number seven, plus or minus two: some limits in our capacity for processing information. Psychol. Rev., 1956, 63, 81-97.

12. Nevel'skij P.B. Sravnitel'noe issledovanie ob'ema kratkovremennoj i dolgovremennoj pamjati. 18 meždunarodyj psixologičeskij kongress. Simpozium 21, 1966, 26.

13. Čistovič L.A., Klaas Ju.A., Alëkin R.O. O značenii imitacii dlja raspoznavanija zvukovyx posledovatel'nostej. Vopr. psixologii, 1961, 5, 173-182.

14. Čistovič L.A., Koževnikov V.A., Aljakrinskij V.V. i dr. Reč'. Artikuljacija i vosprijatie. "Nauka", 1965.

15. Mehler S. Some effects of grammatical transformations on the recall of English sentences. S.verb. Learn.verb. behaviour, 1963, 2, 346-351.

16. Zagorujko N.G. Kombinirovannyj metod prinjatija rešenij. Sb. tr. IM SO AN SSSR. "Vyčislitel'nye sistemy", vyp. 22, Novosibirsk, 1966.

17. Zagorujko N.G. Kakimi rešajuščimi funkcijami pol'zuetsja čelovek? Sb. tr. IM SO AN SSSR. "Vyčislitel'nye sistemy", vyp. 28, Novosibirsk, 1967.

18. Čistovič L.A., Fant G., Serpa-Lejta A., T'ernlund P. Kvartal'nyj otčet laboratorii peredači reči Stokgol'mskogo Korolevskogo texnologičeskogo instituta, No. 4, 1966.

19. Miller G.A. Decision units in the perception of speech. IRE Trans. on Information Theory, 1962, 8, 81-83.

20. Galunov V.I. Nekotorye osobennosti vosprijatija reči. Akust. ž. 1966, 12, 422-427.

21. Vološin G.Ja. Ob ispol'zovanii jazykovoj izbytočnosti dlja povyšenija nadežnosti avtomatičeskogo raspoznavanija rečevyx signalov. Sb. tr. IM SO AN SSSR "Vyčislitel'nye sistemy", vyp. 28, Novosibirsk, 1967.

22. Zagorujko N.G., Lozovskij V.S. Podstrojka pod diktora pri raspoznavanii ograničennogo nabora ustnyx komand. Sb. tr. IM SO AN SSSR, "Vyčislitel'nye sistemy", vyp. 28, Novosibirsk, 1967.

23. Fujisaki H. a. T. Kawashima. The roles of pitch and higher formants in the perception of vowels. 1967 Conference on Speech Communication and Processing, 251-256.

24. Lisenko D.M. Principy vydelenija i morfologičeskogo analiza slova pri vosprijatii ustnoj reči. Diss., L., 1966.

25. Flanagan J.L. Speech, Analysis, Synthesis and Perception. Academic Press, New York, 1965.

26. Molčanov A.P., Labutin V.K. Slux i analiz signalov. Izd. Energija, 1967.

27. Glezer V.D. Mexanizmy opoznanija zritel'nyx obrazov. Nauka, 1966.

28. Al'tman Ja.A., I.A. Vartonjan, G.V. Geršuni, A.M. Maruseva, E.A. Radionova, G.I. Ratnikova. O funkcional'noj klassifikacii po vremennym xarakteristikam nejronov stvolovyx otdelov sluxovoj sistemy. Doklad, pročitannyj na Naučnoj sessii instituta fiziologii im. I.P. Pavlova AN SSSR, posvjaščennoj 50-letiju Velikoj Oktjabr'skoj Socialističeskoj revoljucii. Oktjabr' 1967 g.

29. Hubel D.H. and T.N. Wiesel. Receptive fields of single neurons in the cat's striate cortex. Journ. Physiol., 148,574.

30. Whitfield J.C. and Evans E.F. Responses of auditory cortical neurons to stimuli of changing frequency. J. Neurophysiology, 1965, vol. 28, 655-672.

31. Jakobson R., Fant G., Halle M. Preliminaries to speech analysis. The distinctive features and their correlates. Acoust. Lab. M.I.T. Techn. Rep. 13. Cambridge, Mass., 1952.

32. Bondarko L.V. Differencial'nye priznaki slogov. Otčet lab. eksperimental'noj fonetiki. L., 24, 1967.

33. Bondarko L.V. Struktura sloga i xarakteristiki fonem. Voprosy jazykoznanìja, No. 1, 1967.

34. Varšavskij L.A. Ob avtomatičeskom raspoznavanii reči. Voprosy radioelektroniki. ser. XI, vyp. 1, 1964, 5-22.

35. Paul A.P., House A.S. and Stevens K.N. Automatic reduction of vowel spectra: an analysis-by-synthesis method and its evaluation. J.Acoust.Soc.Am., 1964, 36. 303-308.

36. Čistovič L.A. Psixoakustika i voprosy teorii vosprijatija reči. Raspoznavanie sluxovyx obrazov. "Nauka", Novosibirsk, 1966, 68-168.

37. Šupljakov V.C. O tonal'noj vysote zvukov /s/ i /š/. Sb. "Mexanizmy rečeobrazovanija i vosprijatija složnyx zvukov." 1966, 87-95.

38. Šupljakov V.C. Sluxovoj analiz stacionarnyx šumnyx soglasnyx. Diss. L., 1967.

39. Šupljakov V.C. Akustičeskij priznak vosprijatija mjagkosti stacionarnyx šumnyx soglasnyx. VI Vsesojuznaja akustičeskaja konferencija. M., 1968.

40. Chistovich L.A., Fant G. A. de Serpa-Leitao. Mimicring and perception of synthetic vowels. Part II. Speech Transm. Lab. Quart. Progr. a. Status Report. Stockholm, 1966, 3, 1-3.

41. Small A.M. and Daniloff R.G. Pitch of noise bands. J. Acoust. Soc. Am., 1967, 41. 506-512.

42. Ekdahl A.G. and Boring E.G. The pitch of tonal masses. Am. J. Psychol., 1934, 46, 452-455.

43. Nábělek I., Krútel J., Majerník V. Ein Beitrag zur Bestimmung der Tonhöhe von Bandpassrauschen. III Akust. Konferencia. Budapest, 308-313.

44. Plomp R. The ear as a frequency analyzer. J. Acoust. Soc. Am., 1964, 36, 1628-1638.

45. Flanagan J.L. A difference limen for vowel formant frequency. J. Acoust. Soc. Am., 1955, 27, 613-617.

46. Miller G. A. and Nicely P.E. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am., 1959, 27, 338-352.

47. Čistovič L.A. Vlijanie častotnyx ograničenij na razborčivost' russkix soglasnyx zvukov. Sb. Telefonnaja akustika. L., 1955, 1-2, 35-42.

48. Derkač M.F. Statistika vosprijatija gluxix vzryvnyx i ščelevyx soglasnyx v zavisimosti ot ix dlitel'nosti. Sb. "Voprosy statistiki reči". L., 1958, 40-45.

49. Žukova M.G. Vosprijatie i vosproizvedenie dlitel'nosti sintetičeskogo soglasnogo v sloge tipa SG. 1968. Trudy VI Vsesojuznoj akustičeskoj konferencii.

50. Fant G. Akustičeskaja teorija rečeobrazovanija. "Nauka", 1964.

51. Čistovič L.A. Psixofizičeskie xarakteristiki sluxa. "Inženernaja psixologija". Izd. MGU, 138-158.

52. Flanagan J.L. Audibility of periodic pulses and a model for the threshold. J. Acoust. Soc. Am., 1961, 33, 1540-1549.

53. Tumarkina L.N., N.A. Dubrovskij. Nekotorye osobennosti vosprijatija čelovekom amplitudno-modulirovannyx signalov. Biofizika, 1966, II, v.4, 653-658.

54. Borovičová B. and Maláč V. Towards the basic units of speech from the perception point of view. Proc. Seminar on Speech Production and Perception. Leningrad, 1966. Z. für Phonetik usw. in press.

55. Öhman S.E.G. Numerical model of coarticulation. J. Acoust. Soc. Am., 1967, 41, 310-320.

56. MacNeilage P.F. and De Clerk J.L. On the motor control of coarticulation in CVC monosyllables. 1967. Conference on Speech Communication and Processing. Boston, Mass., C3, 157-163.

57. Sorokin V.N., Fajn V.S. Nepreryvno-gruppovoe raspoznavanie slov: algoritm i eksperimental'nye resul'taty. Trudy VI Vsesojuznoj akustičeskoj konferencii. 1968. M.

58. Stevens K.N., House A.S., Paul A.P. Acoustic description of syllabic nuclei: an interpretation in terms of a dynamic model of articulation. J. Acoust.Soc.Am., 1966, 40, 123-132.

60. Liberman A.M., Delattre P., Cooper F.S. and Gerstman L.J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monographs, 1954, 68, No. 8, p. 1-13.

61. Delattre P., Liberman A.M. and Cooper F.S. Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am., 1955, 27, 768-773.

62. Harris K.S., Hoffman H.S., Liberman A.M., Delattre P., Cooper F.S. Effect of third-formant transitions on the perception of the voiced stop consonants. J. Acoust. Soc. Am., 1958, 30, 122-126.

63. Hoffman H.S. Study of some cues in the perception of the voiced stop consonants. J. Acoust. Soc. Am., 1958, 30, 1035-1941.

64. Öhman S.E.G. Coarticulation in VCV utterances: Spectrographic measurements. J. Acoust. Soc. Am., 1966, 33, 151-168.

65. Stevens K.N. Acoustic correlates of certain consonantal features. 1967 Conference on Speech Communication and Processing. Boston, Mass., C6: 177-184.

66. Rejtblat L.E. Vosprijatie napravlenija izmenenija častoty spektral'nogo maksimuma v sintetičeskom sloge. Diplomnaja rabota, 1968.

67. Liberman A.M., Delattre P.C., Gerstman L.J. and Cooper F.S. Tempo of frequency change as a cue for distinguishing classes of speech sounds. J. exp. Psychol., 1956, 52, 127-137.

68. Brady P.T., House A.S. and Stevens K.N. Perception of sounds characterized by a rapidly changing resonant frequency. J. Acoust. Soc. Am., 1967, 33, 1357-1362.

69. Heinz J.M., Lindblom B.E.F., J.Ch.K-G. Lindquist. Patterns of residual masking for sounds with speech-like characteristics. 1967 Conference on Speech Communication and Processing. Boston, Mass., D1, 246-251.

70. Čistovič L.A. Izmenenie osnovnoj častoty golosa kak različitel'nyj priznak soglasnyx. Akus. Ž. v pečati.

71. Heinz J. and Stevens K.N. On the properties of voiceless fricative consonants. J. Acoust. Soc. Am., 1961, 33, 583-596.

72. Liberman A., Harris K.S., Eimas P., Lisker L., Bastian J. An effect of learning on speech perception: the discrimination of duration of silence with and without phonemic significance. Language and Speech, 1961, 4, 175-195.

73. Huggins A.W.F. How accurately must a speaker time his articulation 1967, Conference on Speech communication and Processing. Boston, Mass., D6, 268-273.

74. Hirsh I.J. Auditory perception of temporal order. J. Acoust. Soc. Am., 1959, 31, 753-767.

75. Čistovič L.A. O procedure raspoznavanija fonem čelovekom. Voprosy psixologii (v pečati).

76. Chistovich L., Fant G., A. de Serpa-Leitao, Tjernlund P. Mimicring of synthetic vowels. Speech Transmission Lab. Quarterly Progress and Status Report, 1966, 2, 1-18.

77. Bondarko L.V. O xaraktere izmenenija formantnogo sostava russkix glasnyx pod vlijaniem mjagkosti sosednix soglasnyx. Uč. zap. LGU, 1960, No. 237, vyp. 40.

78. Bondarko L.V. i Zinder L.R. O nekotoryx differencial'nyx priznakax russkix soglasnyx fonem. Voprosy jazykoznanija. No. 1, 1966.

79. Verbickaja L.A. Zvukovye edinicy russkoj reči i ix sootnošenie s ottenkami i fonemami. Kand. diss., L., 1965.

80. Čistovič L.A. Klassifikacija zvukov reči pri ix bystrom povtorenii. Akust. ž., 1960, 6, 392-398.

81. Aljakrinskij V.V. Imitacija det'mi (4-7 let) russkix i nekotoryx anglijskix glasnyx. Voprosy psixologii, 1963, 4, 80-87.

82. Galunov V.I. Struktura množestva rečevyx obrazov. Diss. L., 1967.

83. Bondarko L.V., Verbickaja L.A., Zinder L.R. i Pavlova L.P. Različaemye zvukovye edinicy russkoj reči. Sb. "Mexanizmy rečeobrazovanija i vosprijatija složnyx zvukov." 1966.

84. N.S. Trubeckoj. Osnovy fonologii. M., 1960.

85. Liberman A.M., Cooper F.S., Studdert-Kennedy M., Harris K.S., D.P. Shankweiler. Some observations on the efficiency of speech sounds. Zeitschrift für Phonetik, 1968, in press.

86. Wickelgren W.A. Distinctive features and errors in short-term memory for English vowels. J. Acoust. Soc. Am., 1965, 38, 583-588.

87. Wickelgren, W.A. Distinctive features and errors in short-term memory for English consonants. J. Acoust. Soc. Am., 1966, 38, 388.

89. Delattre P. From acoustic cues to distinctive features. Rezjume dokladov 6-go Meždunarodnogo kongressa fonetičeskix nauk. Praga 1967, 33.

38

90. Lindner G. Veränderung der Beurteilung synthetischer Vokale unter dem Einfluss des Sukzessivkontrastes. Zeitschrift für Phonetik, 1966, 19, N. 4/5 - 287-307.

91. Kok E.P. Izbiratel'noe rasstrojstvo v organizacii reči: nestojkost' sluxo-rečevyx sledov. Voprosy psixologii. 1965, 2,28-34.

92. Öhman S.E.G. Numerical model of coarticulation. J. Acoust. Soc. Am., 1967, 41 (2), 310-320.

93. Bondarko L.V., Zinder L.R., Pavlova L.P. Različaemye zvukovye tipy russkix soglasnyx. Voprosy radioelektroniki. TPS, vyp. 5, 1967.

94. Bondarko L.V. Vosprijatie differencial'nyx priznakov i slogovaja struktura reči. Tezisy dokladov VI Meždunarodnogo kongressa fonetičeskix nauk, Praga, 1967.

95. Zagorujko N.G. Problema raspoznavanija reči kak složnaja sistema. Tezisy dokladov VI Meždunarodnogo kongressa fonetičeskix nauk. Praga, 1967.

96. Wald A. Statistical decision functions. John Wiley & Sons, N.Y., 1950.

97. Evrejnov E.V., Kosarev Ju. G. Odnorodnye universal'nye vyčislitel'nye sistemy vysokoj proizvoditel'nosti. Izd-vo "Nauka" Sib. otd. Novosibirsk, 1966.