Piotra Łobacz

PROCESSING AND DECODING THE SIGNAL IN SPEECH PERCEPTION

WARSZAWA 1981

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 13 listopada 1980 r.

Zarejestrowana pod nr5/1981

Praca została wykonana w kooperacji z Instytutem Językoznawstwa UAM w ramach problemu węzłowego 10.4 /zlec. 48/

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN

Nakład 140 egz. Ark.wyd. 4,7. Ark.druk. 6 .
Oddano do drukarni w lutym 1981 r.
Nr zamówienia WDN zam. 101/o/81 n.140

Warszawska Drukarnia Naukowa, Warszawa,
ul.Śniadeckich 8

Piotra Łobacz
Institute of Linguistics
Mickiewicz University Poznań

2.23 - rozpoznawanie mowy

The research was carried out under a contract with IFTR within Project No 10.4

PROCESSING AND DECODING THE SIGNAL IN SPEECH PERCEPTION

Abstract

An attempt is made to summarize systematically present-day knowledge of speech processing by man with a view to its inclusion in automatic systems in which word recognition is preceded by phoneme identification. It is suggested that the most interesting models are those which assume (a) parallel processing in several blocks, (b) interaction between bottom-up and top-down procedures, (c) decoding phonemes on the basis of Distinctive Features, (d) categorical character of perception at the segment level, (e) the acceptance of the lexeme as the basic element of recognition. Several theoretical approaches to the models are discussed and the following distinctions between the models are made: (a) hierarchical vs. parallel, (b) passive vs. active. The relation between primary recognition, linguistic processing and the operation of the Precategorical Auditory Storage are expounded. Particular attention is focussed on categorical perception and the theory of feature detectors. Phonemes are shown to have psychological reality. Three lexical access models are surveyed.

1. Theories and models of speech perception.

Speech perception is mostly considered to be a process or a sequence of processes leading to the comprehension of the message the listener has heard. According to Fant (1973) speech perception is a process consisting of both successive and simultaneous identification on a series of progressively more abstract levels of linguistic structure. The consecutive processes are therefore themeselves more and more abstract. Also in Massaro's view (Massaro 1972), perception is a temporal sequence of identifications or recognitions. A similar definition is given by Fujimura (1968) for whom perception is the identification of the signal's form representing a unit, or units, in the descriptive structure of the information transmitted by that signal. Studdert-Kennedy (1973) regards perception as extraction from the acoustic signal of a message coded in accordance with the rules of natural language. In Pisoni's opinion (Pisoni 1975 b) perception is a hierarchically organized sequence of effects related to the storage and the transmission of information. A similar conception is proposed by Foss and Swinney (1973). What these authors mean by perception is a transformation of the input signal into a message in terms of the following stages of analysis: (1) auditory, (2) phonetic, (3) phonological and (4) lexical, syntactic and semantic. Some researchers limit their view of perception to the functioning of the sensory system. Chistovich (1971) maintains that perception is a sensory transformation of a stimulus and the analysis of the result of that transformation. Norman and Rumelhart (1971) include in perception those operations which are involved in the initial conversion of the incoming physical signal into a sensory image, extraction of the relevant features from that image, and identification of an ensemble of features with a previously learned structure. In this conception perception is the first of a series of three successive linguistic stages in the processing of linguistic information. The succeeding ones are: memory and mental decisions (cf. Kurcz 1976). In most of the formulations referred to above, the authors take advantage of the linguistic knowledge and include the functioning of a memory. Cutting and Pisoni (1976) formulated the following hypotheses:

(a) Perception is a process.

(b) There are several stages of signal transformation.

(c) Perception requires storage (memory).

(d) Memory stores have limited capacity and range.

In this paper, we shall discuss the main tendencies in the research work on speech perception and briefly review the most significant results of individual investigations. Particular attention will be given to the perceptual phenomena leading up to, and involved in processing in the speech mode on the level of the phoneme.

Over the last thirty years, a great many experiments have been made to investigate the perception of both segmental and suprasegmental speech units. Fragmentary as these experiments may be, they have been used to form bases for theories and general models of linguistic-information processing by man. It is not a coincidence that the most penetrating theories of natural speech perception should have originated in well-known communications research centres (see Liberman 1961; Liberman et al. 1963, Liberman et al. 1964, Liberman et al. 1967, Liberman et al. 1968 on the motor theory of speech perception; Stevens 1972 a, Stevens and Halle 1967, Halle and Stevens 1959, Halle and Stevens 1964 on Analysis-by-Synthesis, later adpopted by Chomsky and Halle, 1968).

A general inclination may be noted in the relevant literature towards presenting general descriptions of how the speech signal is processed by man, mostly in the form of flow charts referred to as models'. Such models can be found not only in general reviews but also in many specifif papers. The publications by Haggard and his colleagues (Allan and Haggard 1977; Haggard 1974) may be named as examples of this. The main reason for this tendency should be sought in the diversity and ramifications of the issues involved in the natural processing of the speech signal. Every investigation, however fragmentary, is embedded in some general theory, not always exhaustively formulated, which, to greater or lesser degree, it is supposed to validate.

A complete and exhaustive treatise in the form of an interdisciplinary monograph including results obtained by acousticians, audiologists, phoneticians, neurologists, neurolinguists,

psychologists, psycholinguists, and others perhaps dealing with the different aspects of the entire problem of linguistic-information processing by man has not yet been produced. As a consequence of this lacuna the proposed models are largely heuristic and only sections of them have really been based on the results of concrete experimentation. The interdisciplinary character of perception research has been pointed out, among others, by Fant (1968) and by Cohen (1980). Formalization of some of the operations involved in speech perception has also been attempted, e.g., by Bondarko et al. (1968) and by Galunov and Chistovich (1965). A successful formal model of the entire process has not yet been worked out: "Because speech perception is a complex process involving several interrelated components, it has been difficult to formalize all relevant aspects of the process into a coherent model" (Pisoni 1978 pp. 225-226).[1]

T e c h n i c a l models are rare, but one example is the proposal made by Tappert (1968). It simulates the initial stage of language acquisition by infants. Klatt's (1979) LAFS and SCRIBER systems may (partly at least) be regarded as such models. These last-mentioned propositions are of the Speech Understanding System - the SUS - type. Among the Automatic Speech Recognition - ASR - systems, the SUSs come closest to natural perception models, emphasizing, as they do, the correct comprehension of complete utterances rather than the identification of individual words (cf. Moore 1978).

A technical system has also been designed which simulates the working of the human auditory mechanism and this system is used, among other things, for automatic segmentation of the speech signal. The designers of this model maintain that "... the only way to describe human speech perception is to describe not the perception itself but the artificial speech understanding system which is most compatible with the experimental data obtained in speech perception research" (Chistovich 1979, pp. 88-89). Studdert-Kennedy (1979) indicated the advanteges for perception research of the achievements in the area of Speech

[1] For the treatment of the memory, though, such a model h a s been developed, cf. Norman and Rumelhart (1971).

Understanding Systems.

Every model of speech perception includes an input and an output with a number of intermediate processing blocks. Each separate level also has its input(s) and output(s) and a number of transformation rules. The transformation processes are presented, in most models, as speculative hypotheses or as generalizations from rather fragmentary studies. This is so because the perceptual processes are largely inaccessible, in their actual working, to direct exploration (cf. Warren 1976). As a consequence, the experimental work, of necessity, avails itself of the following methods:

(1) Artificial isolation of particular stages of the complete processes limiting the experiment to the study of, e.g., the identification of phonetic segments using monosyllabic nonsense words as stimuli, this usually implying elimination of the lexical and syntactical stages.

(2) Production of artificial stimuli with the aid of speech synthesizers which generate signals with preselected values of chosen parameters. Sure enough, this permits a thorough study of the effects of the chosen factor (such as formant transition) on the perception in a particular processing block, but the price to pay is considerable loss of naturalness of the stimuli.

(3) Application of various kinds of interference and distortion, both as a function of time and frequency (e.g., masking by noise, gating out selected fragments, asynchronous transmission to left and right ear, etc.).

(4) Study of pathological cases (such as hearing loss or aphasia).

(5) Observation of the process of retention and recall of preselected syllables or lexemes.

(6) Study of the acquisition of language by the infant and the development of language in childhood.

In most of such manipulations and observations, the experimenter's interference in the natural course of the perceptual process leads to the concentration of the subjects' attention on the preselected factor, which possibly could otherwise not enter their awareness. The effect of artificial attitudinal concentration is often unjustifiably ignored in the interpretation of

the results.

1.1. General conceptions and the characteristic features of the models of human speech signal processing.

Most of the models proposed in the literature, as indicated above (p. 6) are in the nature of heuristic block diagrams. Actually constructed systems are often referred to as 'functional models'. However, the concept of 'functionality' has not always been conceived in the same way by different authors. Stevens and House (1971), for instance, admit that the model proposed by them can be regarded as no more than functional because they are not in possession of the necessary data on the neurophysiological process. According to Morton (1971), a functional model distinguishes processes in the brain that differ on one more of the following criteria: the kind of code used in information processing, the type of information that can interact, and the logical form of the processing operation. Morton's 'logogen' system is functional in the sense that it emphasizes the differences of coding in the individual blocks.

In the construction of 'functional' diagrams of the entire process one or two non-contradictory assumptions out of the following four have been made:

(1) The system of speech processing is in its entirety hierarchical and serial.

(2) Some operations (or even most of them) are performed in parallel.

(3) The system is passive. The means that the analysis within it consists in matching the input signal (primary or transformed) against stored prototypes.

(4) The system is active (dynamic) as, while the speech signal is being processed, identyfying hypotheses are generated and tested against the input.

All the models largely reflect a linguistic analysis of the message. Firstly, they assume that when perceiving speech, man extracts all, or at least most of the linguistic entites beginning with the allophone and ending with the sentence (as a definite syntactic structure), these being units established by linguists in their study of complete utterances. Further, there is also frequently the assumption that listeners react to the

sound of the received message by appealing to their own linguistic knowledge accumulated in the process of language acquisition. Liberman et al. (1964) claim that the linguistic structure is a description of what is perceived. Marslen-Wilson maintains, "Thus any attempt to understand the process of sentence perception will at some point have to come to grips with the problem of how linguistic knowledge is represented in the mind of the listener" (1976 p.203). Studdert-Kennedy (1980) assumes that the listener as the receiver of linguistic information possesses the knowledge of abstract linguistic entities: "To perceive speech is to apprehend this structure [syntactic and semantic], an act that can only be accomplished by the listener who knows a language. Just what the listener knows is, of course, very much the question. But he must, at least, know the abstract linguistic units that compose the massage - whether words, morphemes or phonemes and must know how such units are realized in the signal" (p.123).

Studies of the natural perceptual system make it possible gradually to discover how in fact man acquires linguistic knowledge. Still, at this time, it is not yet possible to establish with certainty whether the subjective experience of language users is in agreement with objective linguistic descriptions of language phenomena. The amount of psycholinguistic experimentation is still insufficient for that. Besides, the definitions of the basic linguistic units such as the phoneme, the syllable, or the word, are still the object of much controversy (see, e.g., Crystal 1971). And since the individual blocks representing the stages of linguistic-information processing, in most models, carry names that are directly related to such basic units (e.g. 'morphemic analysis', 'lexical analysis', etc. , differences of opinion in linguistic theory are naturally reflected in perception studies.

1.1.2. Series models.

An example of a model entirely in accord with the traditional linguistic analysis is the procedural scheme proposed by Bondarko et al. (1968) and by Chistovich et al. (1968), which appear here as Fig. 1. This is a strictly hierarchical model of both the auditory and the neural processes. Each block deals with the out-

put from the preceding block. Moreover, each successive output is characterized by a smaller number of parameters. As in most other models, the first stage of recognition introduced by these authors, viz. the auditory analysis, is the peripheral processing of the acoustic signal in psychoacoustic terms. There are considerably more psychoacoustic properties than there are abstract phonetic features represented in the signal after linguistic processing.[1] In man-to-man communication, only part of the information contained in the signal is used. This part is referred to by the Soviet authors as u s e f u l information (Galunov and Chistovich 1965). The seemingly formal model proposed by them is a multistage representation of external control by means of speech. It is presented as the chain

$$\omega_i \longrightarrow x_i \longrightarrow y_i \longrightarrow \xi_i^{(1)} \longrightarrow \ldots \xi_i^{(k)} \longrightarrow \Omega_i$$

where

ω_i belongs to the set of sound signals $\{\omega_i\}$,

x_i is an element of the set $\{x_i\}$ representing the first stage of recognition, i.e., a set of auditory sensations,

y_i belongs to the set of representations in the speech mode $\{y_i\}$ and

Ω_i is an element of the set of final products. There is strong progressive reduction of information in the 'chain'.

A hierarchical system of speech processing was also proposed by Liberman (cited in Pisoni 1978). The main feature of this model (see Fig. 2) is its versatility. It relates to all the stages of information processing both in the generation (top-down) and the reception (bottom-up). The design of this model is similar to that of Bondarko et al. (1968) in that each successive level of analysis is used to transform and re-code the acoustical signal into increasingly abstract representations (in the case of perception) or, conversely, in gradually more concrete forms, from concepts to the physical signal, in speech production. But

[1] The Soviet authors do not use the concept of distinctive features in the formulation of Jakobson, Fant and Halle (1952) which they distinguish in the phonetic block are, however, quite close to those in the classical approach, especially in the sense of the motor definitions.

Liberman's model was based on a different linguistic philosophy His theoretical foundation was generative grammar rather than the traditional type of description. A diagram including procedures like those in Fig. 2, with details of the separate analysis levels, transformational rules and signal forms at each level is presented in Lieberman et al. (1967).

1.1.3. Models with partially parallel processing.

The sequential character of information processing in the form of analysis in successive blocks raises a number of questions. "It is often assumed that when we are listening to speech we are making phonemic decisions, syntactic decisions, semantic decisions - all more or less simultaneously. Since the decision units involved in these hypothetical decisions are hierarchically related, one might expect that decisions would be made first at the lowest level, then the outcome would provide a basis for decisions at the next higher level, etc. No doubt this approach is reasonable and could be made the basis for a device to recognize speech, but there are several reasons for doubting if it describes the way people naturally operate" (Miller 1962, p. 81).

With respect to the postulation of the different analysis blocks, the signal-to-message transformation of Studdert-Kennedy (1973) is comparable to that of the Soviet authors except that in the former the semantic and the syntactic stages are combined in one block with the lexical analyzer. Studdert-Kennedy maintains that though the individual stages make up a hierarchical structure, information processing in the model must proceed both in series and in parallel so that decisions at higher levels could affect those at lower levels (see also Studdert-Kennedy 1976a). Several authors point out that a listener cannot, in a very brief time interval, make too many sequential decisions (Lieberman 1967, Klatt 1979). Perceptual analysis and front-end recognition are often treated as being in series (see, e.g., Massaro 1970, Haggard 1974, Pisoni and Sawusch 1975, Massaro and Cohen 1975, Oden and Massaro 1978, Jassem 1977 and 1979). But there are also hypotheses proposing that the processes taking place in the linguistic blocks may affect even the quite peripheral analysis (Fujimura 1968). Cutting and Pisoni (1976) claim that though the auditory analysis, logically speaking, precedes

phonetic processing, in most situations the system functions so that both processes take place in parallel. For Liberman et al. (1968) auditory and phonetic analyses must be simultaneous as, at any point of time the acoustical signal carries information about more than one phoneme. Cole and Scott (1974) created a model in which, at the level of acoustic-phonetic analysis there are t h r e e processors, part integrated and part independent, which require simultaneous identification of three basically different types of phonetic information-bearing entities. The fact that decisions at higher levels (with awareness) may be taken before those at the lower levels was pointed out, e.g., by Darwin (1976) and by Rubin et al. (1976).

Marslen-Wilson is of the opinion that every word heard in the context of normal conversation is directly subjected to simultaneous processing at all levels phonetic, lexical, syntactic and semantic making use of the information available at other levels and at all points along the entire sentence.

The human speech recognition system is extremely effective. It employs two strategies: bottom-up (hierarchical) and top-down ('componential') (cf., especially, Marslen-Wilson 1975 and Marslen-Wilson and Welsh 1978). The bottom-up procedure is also described as being 'data-driven' and the top-down one as being 'conceptually-driven'. This corresponds largely to inductive vs. deductive operatons (cf. Norman 1976). Haggard (1975), following Vinograd, suggests the term 'heterarchical' as opposed to 'hierarchical', in accordance with the following tenet: A process at one level may induce processes at any other level, without hierarchical information transmission. In his own model (Haggard 1974) he exemplifies such 'heterarchy'.

In 1972, Stevens and House (Stevens and House 1972) were still uncertain whether the hearer uses his internalized knowledge about the semantics, the syntax and the phonological rules of the language in order to decode the auditory patterns, whether linearly or in parallel. At present, on the basis of increasingly intense experimentation (Savin 1963, Pickett and Pollack 1963, Galunov 1968, Maruszewski and Nowakowska 1970, Warren 1970, Savin and Bever 1970, Kozhevnikov and Chistovich, cited in Stevens and House 1972, Stevenson 1973, Foss and Swinney 1973, Shockey

and Reddy 1974, Warren 1976, Marslen-Wilson and Welsh 1978), it may be assumed, more confidently than some time ago, that linguistic-information processing by man is, to a great extent, a parallel process.

The models presented below illustrate the theory of parallelism of at least some of the procedures is speech perception. The following different operations are usually postulated: several parallel transformations in one complex block or else processing in several blocks with feedback (information containing errors is referred back to lower levels for re-processing) and feedforward (information in the given block is only partially decoded but is transmitted immediately for further processing in an incomplete form). Fant (1967) submitted a so-called 'auditory' theory of speech perception (a term used in Pisoni's review papers 1970 and 1978). In this theory, Fant introduces the concept of auditory distinctive features understood to be subphonemic auditory patterns whose combinations form phonemes, syllables, words and prosodemes. The model designed according to this theory includes two feedback circuits, one in the periphery, between the receptor and the auditory analysis block in the central nervous system, and the other between the block of final message processing and the block of higher-order linguistic units. It can be inferred from the postulation of a common block for phonemes, syllables, words and prosodemes, that this author assumes segmental and suprasegmental feature extraction to be simultaneous.

In a later publication (1973), Fant proposed a generalized scheme for speech recognition, both natural and technical. This model consists of 5 successive stages of processing (cf. Fig. 3):

(1) acoustic parameter extraction,

(2) microsegment detection,

(3) identification of phonetic elements (phonetic transcription),

(4) identification of sentence structure (identification of words),

(5) semantic interpretation.

Each stage includes stores of previously acquired knowledge. The stores hold inventories of units extracted at the given level

with constraints on their occurrence in messages. The system also includes comparators which are placed between the successive stages of processing.

The model provides for a possibility of direct connection between the lowest and the highest levels. The microsegment extraction block replaces the auditory subphonemic pattern recognizer in the earlier model.

Fant's main aim in conceiving his model was to present hypothetical strategies for automatic speech recognition, with reference to the natural process as auxiliary reasoning and the system was developed in some detail, defining all the possible connections between the elements.

A general model of speech recognition both by man's biological system and by a technical system was also proposed by Jassem (1976, 1977). There are two versions of this model, both presented in the form of fairly detailed schemata. The units of phonetic processing are termed allophones in the earlier version (Jassem 1977) and phones in the more recent one (Jassem 1979, see Fig. 4 here). The phones are defined as classes of segments between which the differences are either random, or systematic with regard to individual voice features. The phones are posited as a consequence of bottom-up strategy assumed for the whole system. A substantive modification was introduced, in the second model, into the grammatical analysis. Now, a simultaneous processing of morphs (lowest-order grammatical and lexical units) and syntactic rules was proposed. Besides processing, in the phonetic blocks, elements of the extent of phones, these models provide for the extraction of phonological distinctive features, which may either replace the phone-processing path or operate in parallel to it. At each of the stages at which information is processed in the speech mode (after normalization), the model is provided with a specialized memory which is treated as a store of an ordered totality of the possible units at that level. At the phone stage there are in fact two stores (or memory blocks): one holds a phone inventory and the other, a set of probabilistic rules defining their occurrence. In Fant's general system, too, each processing stage has a separate store of apriori knowledge which corresponds to the memory blocks in Jassem's

models. From the point of view of automatic speech recognition the assumption of such highly specialized stores seems justified. The functioning of the memory systems in natural perception, where retention, storing and recall of the information are of prime importance, may be distinctly different.

An example of a relatively simple system of information processing including the operation of human memory is Broadbent's filter model (Broadbent 1958). The author's main thesis is that man is able to analyse and identify only a limited amount of the information that reaches his sensory inputs. His system functions as a selective filter which may accept the desired message and ignore everything else. After preliminary sensory analysis, information is held in a temporary store, from where it passes to a selective filter connected to a channel of limited capacity. The filter and the channel belong to the central nervous system. There is a double output from the channel, one to the effectors and the other to long-term memory (which stores the conditional probabilities of the events that have occurred in the message) and then to the terminal output. Information may, however, be returned to the short-term memory (the temporary store).

Another model of auditory processing of linguistic messages employing memory was presented by Massaro and Cohen (1975). Structural and functional components were distinguished here. The functions are - feature detection, primary recognition, higher-level recognition and reharsal with recoding (in case of erroneous recognition). The structures are - the two receptors followed in series by three stores: auditory precategorical, auditory synthetic and generated abstract. This model has two properties reminiscent of the early proposal of Fant (1967). First, the synthetic auditory memory is a complex block including several sub-blocks in which information is stored in parallel. In Fant's model the second last analysis block (for higher-order linguistic units) included a set of parallel procedures. Further, in both models there is feedback between the last two operations. Massaro and Cohen propose feedback between meaning and second-level recognition via the reharsal-and-recoding block. A distinction between preliminary and higher-level recognition is in agreement with Fry's conception (Fry 1970).

Massaro and Cohen's model was later modified by Oden and

Massaro (1978) and is reproduced here in Fig. 5. The authors also developed an algebraic form of the model using fuzzy logic in the identification of the phonemes. The functions and the corresponding processing blocks are the same as in the earlier model reviewed above. But a block of long-term memory was now introduced with connections to the recognition block and the rehearsal-and-recoding block. The features in the precategorical auditory store are the direct result of both the auditory stimulus and the properties of the receptors. The detection of the features is not modified by the listener's prior knowledge. During preliminary recognition, every acoustic property is evaluated in the precategorical auditory store and compared with prototypical perceptual units in the long-term memory. Recognition consists in finding the prototype which best matches the feature in the auditory store. The perceptual unit in the long-term memory is very flexible as it is being affected by a number of normalizing algorithms. These effects may be somewhat reduced in the perception of continuous speech in view of contextual contraints and redundancy top-down strategy. One of these authors Massaro 1970 and 1972) studied the functions of the precategorical auditory store and found that the perceptual unit is of the extent of a syllable and that its duration is of the order of 250 ms. The output from the primary recognition block is held in the synthetic auditory memory. Secondary recognition acts on the perceptual information and superimposes meaning on it. It "translates" the perceptual code of the auditory storage into an abstract memory code (cf. also Massaro 1979). Long term memory is treated in this model as an apriori-knowledge source which is not only semantic organized knowledge source about the language, its use and the outside world . The author also admits the possibility of storing, in long-term memory, elements of lower order than the lexicon and the syntax. The synthetic auditory memory plus the generated abstract memory seems to correspond to what other authors refer to as the so-called dynamic short-term memory (Shiffrin and Atkinson 1969, Forrin and Morin 1969, Cutting and Pisoni 1976).

There is no agreement among specialists studying problems of

memory in relation to language as to the definitions and the functions of the different kinds of memory. At the beginning of this paper (p. 4), a few selected interpretations of perception were given. Among them, there are some that treat information processing as a preliminary to memory. This is, for instance, the case in Norman and Rumelhart (1971) quoted above (p. 11). These authors suggested separate mechanisms for sensation, perception, short-term memory, long-term memory and recall. Kurcz (1976) made similar assumptions for her "scheme of information processing by man including linguistic processes" (p. 217). In this scheme, information received from the visual and auditory receptors is held temporarily in the respective sensory stores: iconic or verbal for the visual receptor and acoustic or phonemic for the auditory receptor. In this model, the next stage - the process of perceptual analysis - involves, in addition to the perceptual analyser, a linguistic analyser, which operates at the phonemic, the syntactic and the semantic level. The sensory store can be regarded as equivalent to what Norman (1969) calls "very short memory". This author maintains elsewhere (Norman 1972) that the different kinds of memory: sensory, precategorical, short-term and long-term, do not exist in isolation. They are related to the process of recognition of incoming patterns. This implies that the memories co-operate in spite of their different mechanisms.

Differences of interpretation and definition of the various stages of biological memory in general persist within the restricted area of processing the acoustic form of language. Controversies are particularly noticeable in the different views on the existence of a separate precategorical auditory store and its properties (e.g. Norman 1972, Crowder 1972 and 1973, Cole and Scott 1974, Studdert-Kennedy 1976a), the functions of the short-term memory (Norman 1969, Shiffrin and Atkinson 1969, Forrin and Morin 1969, Wickelgren 1969a, Norman and Rumelhart 1971, Massaro 1974) and also on the structure of semantic (long-term) memory.

In Morton's paper "A functional model for memory" (1971), a model of word recognition is presented (Fig. 6 here) in which only one kind of memory is postulated. It is called "the cognitive

system". Other parts of the model also store information temporarily in greater or lesser amounts, but their function is here assumed to be marginal. Morton maintains that alleged differences between short-term and long-term memory are in fact little more than differences in time spans between the presentation of the stimulus and the response and that they do not involve different modes of storage. The so-called "logogen" model proposed in that paper is an improvement on an earlier version (Morton 1969) and a still earlier one developed in co-operation with Broadbent (Morton and Broadbent 1967). The earliest, most detailed version of the logogen model was worked out in 1964 (according to Morton 1969). In the 1967 system, visual and auditory information reaches the logogen block via an equivalent of an analog-to-digital converter. It is then directed to the memory, but the operations on the logogens - approximately, psychological correlates of words, or morphemes - are also dependent on information from an additional source, viz. the context, in the form of higher-order units, with their semantic implications. The 1969 model separated the two kinds of analysis: auditory and visual, introduced an output buffer block as well as a feedback involving the scanning and rehearsal of the information contained in the buffer. If the response is incompatible (assumed wrong), the information is returned to the logogen block (Morton 1969). The 1971 version (Morton 1971), reproduced here as Fig.6, replaces the context block by a 'cognitive system' with continuous feedback with the logogen block. The output from the complete system - the response - depends on two feedback processes: silent rehearsal which only engages the buffer and the logogen block, and articulated rehearsal which also engages the auditory analysis. The second feedback loop is only activated under special circumstances, for instance when a detailed transcription of the message is needed or when the signal-to-noise ratio becomes very low or when the information is unusually tightly packed in the signal. In all Morton's models, the central block is the logogen system which works on the principle of a c o u n t e r .

The count increases as a function of several inputs until it reaches a treshold value. When this is exceeded, the word corresponding to the activated logogen is recognized. The logogen is p a s s i v e in the sense that its working does not require any (outside) mechanism to perform the matching. Only decisions made inside the logogen block are used.

The common feature of all models reviewed in this sub-section (at least some of them are partially parallel) is that they are in fact passive (cf. point 3 above p.7). Two general procedures are usually employed in such models. One of these is m a t c h i n g against prototypes, i.e., standarized patterns of the given category: "Perception or perceptual processing can be thought of as an analysis of the sensory input in which features are detected that correspond to encoded features in memory. When a sufficient number of features are found in the sensory input that correspond to the features of the item in memory, the stimulus is recognized as that item" (Massaro 1970, p.411). The other procedure is filtering as represented in Broadbent's model (see above).

According to Glucksberg and Danks (1975), Denes (1965), liberman et al. (1967)and Pisoni and Sawusch(1975), models which use the above two procedures are inappropriate. Specifically, Pisoni and Sawusch (1975) assert that segments cannot be defined exclusively by the physical properties of the signal. In Studdert-Kennedy's view (1976a) models based exclusively on prototype matching are inadequate because of the excessive acoustic variability of phonetic segments, due to the effect of context, rate of delivery, accent, individual voice features, etc. This author concedes, however, that a certain amount of matching does occur.

1.1.4. Active models.

In order to show how a c t i v e models work, we have chosen representative systems which have arisen from the following three theories of linguistic-information processing:

(1) motor theory of speech perception,

(2) analysis by synthesis,

(3) lexical access.

The basis for the design of active models is the assumption

that the hearer acts not only when he produces a linguistic message, but also when he receives it - in the latter case by matching his internal speech in the form of generated hypotheses against the speech signal that reaches his auditory receptor (cf. Stevens and Halle 1967, Stevens and House 1972, Cooper 1972).

The main theses of the first of the three theories (also called the "theory of motor commands", cf. Lieberman at al. 1962) are contained in the following quotations: "Articulatory movements and their sensory effects mediate between the acoustic stimulus and the event we call perception" (Lieberman et al. 1967, p. 122). "...proprioceptive as well as acoustic stimulus dimension must be taken into account in any attempt to explain the perception of speech" (Liberman 1971, p.149). "Of all the speech events - -acoustic signal, articulatory shape or neuromotor commands - - about which we can reasonably expect to collect information, the neural commands to the articulators will in our view provide the simplest relationship to phoneme perception (Liberman et al. 1962, p. 8). Liberman and his associaties, the initiators of this theory, assert that the invariance of phonemic entities is much more evident at the motor than at the acoustic level (see also Chistovich 1971, Wickelgren 1969b). In his central nervous system, the hearer is therefore assumed to compare the incoming signal with classes of sounds which he knows from his own a r t i c u l a t o r y experience, even though in the process of perception he does not necessarily activate his articulatory organs (Cooper et al. 1958). Lieberman et al. emphasize that speech is a special kind of code with a special perceptual mechanism which involves specific processes that have access to motor mechanisms. The device that decodes the signal at the phonetic level is peculiar to man as part of his species-bound linguistic ability (cf. Studdert-Kennedy 1970). Fig. 7 here represents, after Liberman et al. (1968), a scheme of the procedures used by the speaker-listener. The hatched area includes the neural interaction between perception and articulation. Information may, in this system, also pass from the auditory receptor to the central nervous system directly, independent of the analysis by the auditory decoder (dashed line in the Figure).

The motor theory of speech perception was, in the sixties,

the subject of heated discussions. Fant (1973) pointed out that there was no apparent advantage in attempting to avoid looking for auditory features of the signal and proposing instead its articulatory reconstruction. In his view (1967) sensori-motor relations may be more necessary at higher levels of processing (in the analysis of sound patterns of complete words). One of the strongest opponents of the motor theory was Lane (1965, 1968 a, 1968 b, 1970). This author repeatedly polemicized with the proponents of the articulatory mediation in perception. In his 1965 paper he gave three conditions which would have to be fulfilled for the motor theory to be considered adequate: (a) Different perceptual responses result from similar acoustic signals if these are produced by different articulatory configurations. (b) Similar perceptual sensations result from different acoustic signals if these are produced by similar configurations of the vocal tract. (c) There is a continuity of variation of the articulatory parameters and a continuity of variation in the acoustic parameters. The variation in p e r c e p t u a l parameters is more closely related to the former that to the latter. Performing his own experiments as well as replicating those of other authors, Lane attempted to show that not all of the relations stated in the conditions actually obtain.

According to Jakobson (1968), the motor feedback is not at all a necessary condition for the identification and discrimination of a verbal message. Denes (1965 and 1967) performed experiments aimed at finding "how far being able to listen to one's own voice, and hereby getting a chance of associating our articulatory movements with the sounds produced by their movements, make learning to recognize speech easier" (1967, p. 310). The results of the experiments neither supported nor refuted the motor theory. Darwin (1976) gives the following kinds of arguments in favour of the independence of speech production and perception: simultaneous translation, results of medical examinations and the failure of the theory to account for the perception of inter-speaker differences. Galunov and Chistovich (1965) produced some experimental evidence in favour of the theory, namely: (1) When recognizing the speech signal in terms of the different vowels and consonants, the hearer places the phonetic elements in a

space whose co-ordinates are motor parameters. (2) with only minimal delay, man can imitate the speech signal he has just heard, which indicates that there is very fast transmition from auditory to motor features. But these authors also give some counterarguments: (1) Children understand spoken language before they can speek, (2) Congentially deaf persons can understand speech, (3) successive proprioceptive images of the articulatory movements can only disappear successively after these movements have been realized by the hearer (which they are not).

At the present moment, the motor interpretation of speech processing has lost much of its topicality and requires modification (Palermo 1975).

K. Stevens, the founder of the theory of a n a l y s i s b y s y n t h e s i s maintains (1972a) that the neural commands are not invariant enough to be able to form reference points for phonetic decoding. The analysis-by-synthesis matching procedure deals with temporal changes of the acoustical energy spectrum according to the assumed quantal nature of speech (Stevens 1971 and 1972b). The prototypes are internally generated in the analyser on instructions which are issued until the best match with the input signal is obtained (Halle and Stevens 1964, Stevens and Halle 1967). The fundamentals of the theory were formulated towards the end of the fifties, together with M.Halle. One of the earliest models (Stevens 1960) appears here in Fig. 8. Its conception leant more on techical criteria than those of natural perception. The block called "Model I" stores rules relating articulatory descriptions to the different speech spectra and "Model II" contains rules which convert phonetic symbols into articulatory descriptions. This second block also serves the purpose of generating the output from the phonetic-symbol analyser. The original system includes two control units and a later version (Halle and Stevens 1964), also two analysis-by-synthesis loops, as shown in Fig. 9. In this version, only one control element was used. In one of the later models (Stevens 1972 a) the control system is a central processing block connected to the preliminary-analysis block, the rule-generating block, the comparator which matches the effects of the rules with the peripheral analysis , the lexicon and the store of the results of previous analysis.

The information about the acoustic properties of the signal is subjected to an initial analysis, after which, in the form of a gross matrix of phonetic features and rules, it is conveyed to the control system, whence it passes directly to the output terminal of the system or, if the control unit gives the proper command, a hypothetical definition of the utterance in terms of morphemes and strings of morphemes is formed. Such information passes on to the block of generative rules. These act on the abstract feature matrix (common for articulation and perception), producing representations in the form prototypical phonetic segments. The comparator, provided with a catalogue of relations between prototypes of the peripheral auditory analysis and the articulatory descriptions, matches the result of the analysis in the block of generative rules with the auditory information previously deposited in the temporary store. The result of the matching procedure returns to the control unit.

Another version of the analysis-by-synthesis theory was proposed by Chomsky and Halle (1968). It was described, much later, by Pisoni (1978) as being innovating. In this theory, the process of speech recognition by the hearer is conceived as follows: "The hearer makes use of certain cues and certain expectations to determine the syntactic structure and semantic content of an utterance. Given a hypothesis as to its syntactic structure - in particular its surface structure - he uses the phonological principles that he controls to determine a phonetic shape. The hypothesis will then be accepted if it is not too radically at variance with the acoustic material, where the range of permitted discrepancy may vary widely with conditions and many individual factors. Given acceptance of such a hypothesis, what the hearer 'hears' is what is internally generated by the rules. That is, he will 'hear' the phonetic shape determined by the postulated syntactic structure and the internalized rules" (Chomsky and Halle 1968). The authors' assumptions were represented by Pisoni (1978) in the form of a tentative scheme here shown in Fig. 11. After having been processed in the block of preliminary analysis, the sensory information passes to the recognition device, whose output is a gross and tentative classification of features and segments. In the phonological block,

decoding rules are used together with syntactic and semantic information in order to obtain lexical representation. The analysis-by-synthesis takes place at the phonological-syntactic level rather than, as in the classical form of the theory, at the neuroacoustic level. The phonological information is used to form hypotheses about the structure of the sentences.

The most recent version of the analysis-by-synthesis system was presented by Klatt (1979). His model, described by its author as being 'simplified' is shown here in Fig. 10. The process of analysis-by-synthesis is now reduced to one block, whilst several memory systems have been distinguished. The first of these, called 'echoic' memory holds informations for about 200-300 ms so that some transitional information may be lost. What is retained is, at any rate, the F_0 contour, the intensity of the envelope, segment duration and vowel quality. The second stage of the analysis is a block of phonetic feature detectors with an output in the form of a feature matrix. This matrix is incomplete because it has been impossible to incorporate in it various contextual conditions. The information obtained from the phonetic feature matrix is used to search the lexicon for word candidates to be sent to the lexical hypothesis memory. The final choice of a word is based on the syntactic-semantic analysis. The analysis-by-synthesis accepts both the bottom-up and the top-down information. Whilst accepting lexical hypotheses, it returns to the acoustic-phonetic data for a check-up of the details of the entire expected word. After syntactic-semantic processing, information returns to the lexical block. This is the information about the words expected (or predicted) in the utterance.

In the models of K. Stevens and his associates the emphasis lay on the phonetic processing of the information. In Klatt's model and that of Chomsky and Halle, the analysis-by-synthesis also included higher-order linguistic units. But an active model which particularly concentrates on w o r d recognition is Marslen-Wilson and Welsh's 1978 d i r e c t l e x i c a l a c c e s s system. These authors start with an assumption, backed by experimental evidence, that it is possible directly to control the lexical interpretation of the phonetic input. The perceptual analysis is followed by acoustic-phonetic proces-

sing. After the first two or three phonemes have been decoded, a whole class of word hypotheses is activated proposing a 'cohort' of word candidates, i.e., the acoustic-phonetic information is passed simultaneously to a number of lexical items in the memory. Each element in the lexical memory is an active processing unit generating prototypical word candidates. The whole system is therefore decentralized. The lexical elements receive information about the contextual conditions (top-down procedure), which determines the final choice among the word candidates at the particular place in the utterance. A rejection causes instant deactivation. The number of activated word candidates thus becomes progresively smaller until the proper (best-fitting) item is found. The principle of choice by selective activation plus elimination by top-down procedure seems more effective than Morton's passive 'counting' model, which works less restrictively.

Such factors as predicting elements of various orders, active utilization of contextual constraints, selection at various stages by correction and rejection increase the speed and overall accuracy of identification, which favours the hypothesis on active processing of linguistic information by man. But a differentiation of active and passive models is not simple, as pointed out quite a long time ago by Licklider (1952). In fact, MacKay (1968) argues that the antithesis of the two theories is unfounded and that the most appropriate model is one which combines filtering procedures with active matching (cf. also MacKay 1967). A compromise solution is proposed by Summerfield and Haggard (1975) with respect to normalization. As mentioned earlier (p. 17), Morton claims that an internal generator only works as an additional element of the system in unusual conditions of perception. According to Cohen (1967) both mechanisms must be employed in any model of speech perception. In his view, in language acquisition the active mechanism may predominate whilst filtering more adequately explains the basic processes of perception in the adult.

2. Perceptual acoustic-phonetic analysis at the word level.

Irrespective of the basic theoretical position, all researchers regard two components of their models as being indispensable:

a primary-analysis block and a grammatical-semantic block. Most of them also assume at least one phonological block.

In the present Section the subject of a somewhat more detailed discussion is,

(1) within the b o t t o m - u p procedure,
   (a) the final result of the primary analysis as a hypothetical input to the phonological block,
   (b) processing in the phonological block and the output of that block,
(2) in the t o p - d o w n procedure,
   (a) processing in the lexical block,
   (b) the effect of feedback between the lexical and the phonological block.

In the various stages, the following kinds of cues are used in the process of decoding the speech signal: acoustic, motor, grammatical, probabilistic and other cues learned during language acquisition (see Singh 1966). All the stages of the perception process are closely related to each other so that even an artificial limitation to the description of the functioning of just some of them cannot obviate frequent references to the others. However, one process has here been left out of the discussion, viz. the normalization of the signal for personal voice features or adaptation of the system to those features.

2.1. Primary processing.

The processes of an initial, or primary, analysis are usually distinguished from the remaining stages of processing by one of four distinctions. Three of these are related to the processes themselves and one to memory. The distinctions are as follows:

(1) peripheral analysis vs. processing in the central nervous system,
(2) auditory analysis vs. processing in the speech mode,
(3) retention of information in the sensory store or a precategorical auditory store vs. short-term memory,
(4) primary recognition vs. linguistic processing.

Each of these distinctions locates the division at a different moment of the entire process. Peripheral analysis is understood to include the function of the outer ear and the cochlear processing. Both are entirely automatic, i.e., extraneous to awareness

(Studdert-Kennedy 1973). Fourcin (1972) distinguishes two stages of the peripheral analysis: the acoustical input to the meatus and the mechanical input to the inner ear. In Fant's model (1967) and in the one proposed by Massaro and Cohen (1975), there are two blocks for the auditory receptors. Haggard's model (1974) includes a rudimentary peripheral analysis in the cochlea. The peripheral-analysis block appears in all analysis-by-synthesis models, in which it is referred to variously as a 'spectrum analyser' (Halle and Stevens 1964), 'spectral analysis block' (Klatt 1979), 'bank of analysis filters' (Stevens 1960) or simply 'peripheral auditory analysis', equivalent to the mechanism of hearing (Stevens and House 1972). Lieberman et al. (1968) describe the first stage of their model as the frequency-time-intensity analysis of the ear. An analysis of the output from the auditory receptors is not the immediate object of the specific description of speech decoding processes. The conversion of mechano-acoustic events into electro-mechanical events is the same for all the sounds received by man. The general properties of the acoustic signal referred to as sensations are the object of audiometry and psychoacoustics and as such they are outside the scope of the present work.

2.1.1. Auditory analysis.

In Jassem's models (1977 and 1979) the peripheral block corresponds to the auditory level in other systems. The difference is, however, not purely terminological because the auditory system, in the interpretation of most authors, also includes processes in the central nervous system. It is being assumed in the present work that the interpretation of the acoustic signal is performed beyond the hearing mechanism along afferent paths and in the subcortial and the cortical hearing centres.

In the models which do not distinguish a separate peripheral block, the auditory receptors are included in the auditory analysis block. The auditory level is regarded as a series of processes. According to Searle, Jakobson and Rayment (1979), three stages of the auditory reception of the signal should be distinguished: filter detection (this corresponds to the peripheral analysis), feature detection (e.g., spectral energy peaks, voice onset time, etc.) and decision making on the output of the feature detectors. According to Pisoni (1973) auditory percep-

tion consists in an analysis of the sound wave as a set of time-varying psychological parameters. It is a transformation of fundamental frequency (F_0), overall intensity, duration, onset and decay amplitude and energy spectrum into pitch, sound quality, loudness, sound type, etc. (cf. also Jassem, 1977). The interpretation of such parameters is performed on the basis of neural patterns. According to Studdert-Kennedy (1973), the closest symbolic correlates of the neural patterns are conventional 'three dimensional' spectrograms. Divenyi (1979) suggests that the speech signal is interpreted by the auditory system in terms of a certain number of discrete psychoacoustic processes. The result of this analysis is in the form of neural spectra, i.e., a series of quasi-stationary auditory events of variable duration. Stevens and House (1972) also assume a transformation of the signal into a kind of neural spatiotemporal pattern, though they consider the nature of such patterns to be still unknown.

2.1.2. Primary segmentation.

During auditory processing, preliminary segmentation is performed. Summerfield and Bailey (1979) assume that a sequence of acoustic events must be first 'segregated' and only then 'detected'. Tsemel (1975) points out that the auditory system reacts well to sudden changes in the overall intensity of the signal and supposes that the mechanism of segmentation reacts to such changes. According to Blumstein et al. (1979) the hearer samples the signal in the vicinity of those moments at which there are rapid changes in the spectrum and classifies the sounds in relation to those moments in terms of certain gross features of the instantaneous spectrum. According to Fant (1973) rapid changes in the spectrum and the intensity introduce quasi-discontinuities of the signal resulting in microacoustic phonetic features. Studdert-Kennedy (1979) maintains that if segmentation takes place at the auditory level at all, it is syllabic rather than allophonic. This view agrees with Massaro's (1972), who believes that fluctuations of acoustic pressure are stored auditorily and organize the incoming sounds in units of the extent of a syllable. In Stevens and House's system (1972) there is partial decoding of the signal and consequently its discretization in an early phase of auditory processing. The scanning of the signal in the peripheral decoder takes place in time intervals close to

the duration of a syllable. In fact the syllable is believed to be the perceptual unit by some authors cf. Liberman et al. 1967, Bondarko 1969, Stevens 1973, Cole and Scott 1974, Studdert-Kennedy 1975). The main arguments are as follows: an unsegmented syllable is the rhythm unit of nursery rhymes, i.e., the earliest unit isolated in language acquisition. The syllable may be determined at a processing level not specialized for speech. It reduces the number of auditory segments emitted per unit time and introduces contrasts in the signal that enable the hearer to discover the individual phone-length segments. For Studdert--Kennedy (1973) the linguistic function of the syllable is to provide the auditory receptors with a rhythm signal which distributes power in time and, by its internal contrast, faciliates the acquisition and development of auditory ability of the hearer to discriminate phonetic segments. According to Bever (1970), the syllable is a relevant unit of speech production and perception. His arguments are as follows: The syllable may be defined relatively easily in acoustic terms, it is the smallest unit pronounceable in isolation and its features are relatively invariant.

When discusing the above properties of the syllable, the authors consider its so-called canonical form, i.e., the sequence consonant plus vowel. Studdert-Kennedy (1976b) claims that such a form of the syllable occurs in all languages. It is therefore a universal phonetic entity. The time parameter is also sometimes a consideration in the definition of the syllable as a perceptual unit. The basic time window for perception is an interval of some 200-250 ms. Lehiste (1972), starting from premises similar to those used by the proponents of the motor theory (mainly that one physical signal such as the formant transition carries information about at least two segments), finds herself compelled to leave the problem of segmentation open. She admits two possibilities: to accept as a minimal perceptual entity either a unit of a lower order than the phone or else the syllable.

In linguistics, the syllable belongs to the most controversial concepts. There are a great number of definitions of the syllable which lead to the conclusion that the division of an utterance

into syllables is extremely dificult to perform with rigorus criteria even though the number of syllables in an utterance is intuitively obvious to every language user (in a vast majority of cases at least). The difficulties in the choice of an adequate definition are particularly apparent in attempts to segment speech automatically. For the purposes of a particular technical procedure, new definitions are usually evolved. Mermelstein (1975) replaced the traditional syllable by a "syllabic unit" because his system of automatic discretization divided the signal into such fragments that syllable boundaries could occur inside phones. A hierarchical procedure of segmentation first into syllables and then into phonemes was developed by Gresser and Mercier (1975). The definition adopted by these authors is: the syllable is a sequence of speech samples containing one vocalic nucleus. However, in order to divide the signal into syllables they need a preliminary segmentation into vocalic, nonvocalic and undefined fragments, so the hierarchy is not particularly consistent. In the second stage of the procedure nine different acoustic-phonetic parameters are analysed to obtain a division into syllables. In the third stage, the successive phonemic units are delimited. The method applied by these authors, then, indicates that vowels have to be located in the speech signal b e f o r e any further automatic segmentation can be performed, after all.

2.1.3. Precategorical Auditory Storage (PAS).

In the description of the working of the auditory block, two aspects mostly come to the forefront: the conversion of the sensory input into a finite number of specific psychophysical parameters, and the time needed for this process, the latter being studied by testing the retention of the output from the auditory block.

The psychoacoustic features are held in the sensory image store for a very short time of about 50 to 250 ms. This store works as a time window shifted along the signal and integrating over approximately 100 ms (cf. Cutting and Pisoni 1976, Pisoni 1978).

The sensory image store SIS is not specialized and it collects information from all receptors. The memory which holds information received by the ear is the Precategorical Auditory Storage, PAS. Massaro and Cohen (1975) start off with PAS. Their model does not include the unspecified sensory store. PAS here acts directly on

the information collected by the receptors and is part of the central nervous system (cf. also Massaro 1974 and Norman 1972). Different features require different detection times. The store accumulates the psychoacoustic features until the sound pattern is completed. Crowder (1972) maintains that PAS retains information about pitch, voice quality, sound quality, sound location and loudness. Material on which the listener's attention has been concentrated is stored differently from signals received without attention. The stimulus, according to this author, is held in PAS for about 200 to 300 ms. It has been found experimentally that PAS stores significantly more information about vowels than it does about consonants (cf., e.g., Crowder 1972, Massaro 1974, Pisoni 1975a) because it responds more strongly to information about stationary events than to dynamic transitional information. It was also shown (Darwin 1975) that the acoustic memory is not dependent on categorization processes because it is an analog representation of the stimuli. In his model, Haggard (1974) introduced a memory system which he called 'intermediate holding register'. The stimulus stored there is defined in terms of relevant acoustic properties. The 'register' is part of the central nervous system, but it is not specialized for speech. In other models, PAS is like Haggard's holding register, the last stage of perception before the speech mode. The fact that the information stored in PAS makes a very gross differentiation of speech sounds possible (by holding information about vowels)[1] characterizes it as an intermediate stage between overall auditory analysis and processing in the speech mode.

The term 'preliminary recognition' mentioned in the last dichotomy on p. 25 mostly refers to all the functions - including storage - related to p r e c a t e g o r i c a l processing. According to Fry (1970) it includes all processes directly dependent on acoustic cues. Massaro (1979) submits that preliminary recognition includes an evaluation and an integration of acoustic features. Haggard (1975) covers by this term the totality of the

[1] It should probably be assumed that PAS stores not only strictly vocalic information, but also information about all vowel-like sounds. The preliminary segmentation performed in PAS may be very inaccurate.

functions of prelinguistic processing.

In conclusion, although there is experimental evidence that purely auditory processes can be distinguished from phonetic processes in some situations of perception (e.g., in studies of dichotic listening, cf. Studdert-Kennedy 1976a), it has also been shown that there are some elements of linguistic analysis in the former also. A sharp line of division between the two stages cannot be drawn, at least not in the light of present-day knowledge. If such a division is assumed, three particularly controversial problems appear: hemispheric specialization, categorical perception, and the so-called feature detector theory.

2.1.4. Cerebral lateralization.

It has been demonstrated in many experimental studies that both in the generation and the perception of speech the activity of the two cerebral hemispheres is not the same. In most situations and cases linguistic processing is more strongly dependent on the functioning of the l e f t hemisphere (cf. e.g., Maruszewski 1970). It has been found (e.g., Bondarko et al. 1968, Pisoni 1971a) that both hemispheres are able to extract complex auditory features such as rapid changes of fundamental frequency, duration, loudness, spectral discontinuity, etc. The signal picked up by the auditory receptors is transmitted to the cerebral hemispheres either ipsilaterally or contralaterally. In experiments on dichotic listening contralateral communication was found. Right ear advantage (REA) was usually shown and it was demonstrated experimentally that, generally, speech or speech-like stimuli were processed more accurately by the right than by the left ear, which is better suited to the analysis of stationary signals. Studdert-Kennedy and Shankweiler, in a paper devoted to lateral specialization in perception (1970), attempted to discover whether speech sounds are processed essentially in the linguistically dominant hemisphere, or in both, to about the same extent. The question was also raised as to whether the signal from the non-dominant hemisphere is transmitted to the specialized hemisphere in the form of an auditory or a linguistic code. The main results of the investigation indicated that formant movements, durational cues, and the extraction of the more general psychoacoustic properties of the speech signal

were processed in both hemispheres. The dominant one (usually the left) is responsible for a more strictly linguistic processing of the signal such as the interpretation of formant transitions in terms of the place of articulation, the extraction of the voicing-feature information, etc. Repp (1975) considered the following three possibilities:

(a) the message, in auditory code, first received by the right hemisphere from the left ear, is then sent to the left hemisphere for the extraction of linguistic information;

(b) phonetic features are extracted in both hemispheres, and a list of such features is sent to the unilateral processor which unites them into phonemes and other units;

(c) the non-dominant hemisphere can decode spoken information, but verbal reactions to the received signal are only generated in the dominant hemisphere. According to this author, the syllable is processed separately in each hemisphere and independently categorized at the subconscious level. The two hemispheres interact before, during, and after the processing of the signal.

It was demonstrated experimentally (Fujimura 1968, Liberman et al. 1967, and, especially, Studdert-Kennedy and Shankweiler 1978) that vowels are distinctly less lateralized, i.e., they are processed to almost the same extent in both hemispheres. They do not engage the discrete-feature extractors in the dominant hemisphere.

The above, of necessity extremely sketchy discussion of the studies of cerebral lateralization leads to the conclusion that linguistic processing begins when the decoding function is taken over by the dominant hemisphere. Liberman (1979) submits, with references to various experimental work, that a given signal may be perceived in two phenomenologically different ways: as speech or as a non-linguistc event. Information about the phonetic segments is contained in various clues. The speech processor sorts these cues and assigns each to the appropriate phonetic construct. The integration of all necessary cues into a single percept is completed at the phone level, the phone being regarded as a linguistic entity. The primary-recognition block in the Massaro-Cohen model (Massaro and Cohen 1975) performs similar functions to those of Liberman's speech processor. According to Summerfield

and Bailey (1979) the 'speechlikeness' is contained in special acoustic attributes of the signal (such as rapid spectral changes, voice onset time, etc.), which - if detected in the initial stage of the auditory analysis - direct the signal to be further processed in phonetic terms.

Under such assumptions, analysis in the speech mode would involve decoding only some consonantal elements of the signal. Vowels and probably also at least the liquid consonants[1] would get decoded before the signal reaches the dominant hemisphere, i.e., at a relatively low level of processing. This assumption, as well as the hypothesis on a Precategorical Auditory Store discussed earlier on, impose an 'automatic' classification into vowel-like and consonant-like speech sounds.

2.1.5. Vowel and consonant.

There are a number of reasons for assuming some basic differences in the way vowels and consonants are perceived. Consonants require longer processing times, are more susceptible to masking effects (Pisoni 1972, Pisoni and Tash 1974). They are discriminated differently from the vowels and they have different access to the short-term memory (Pisoni 1973). There are also differences in the order of processing segmental and suprasegmental feature: Vowels may be processed independently of the suprasegmentals whilst consonants are always decoded after a suprasegmental analysis (Miller, 1978). According to Studdert-Kennedy (1976a), the duration of a vowel is perceptually redundant from the point of view of the time needed for its identification: Correct recognition may be obtained on presentation of no more than one or two periods of a vowel sound (see also Huggins 1964). This is not enough for the identification of a voiced consonant (Matuszkina and Mikiel

[1] A decided majority of the experimental studies investigating the perception of vowels and consonants have been carried out on a few selected vocalic phonemes, usually /a/ and /e/, and the stop consonants, usually the voiced, or lennis, /b d g/. General conclusions drawn from a comparison of only such highly selective stimuli seem premature.

1979, Krawiec et al. 1978).[1]

The perception of consonants is more strongly related to the linguistic categories of phonetic processing (Stevens and House 1972). Fant (1973) postulates separate distinctive features for vowels and for consonants. He polemicizes with Chomsky and Halle (1968), who assume a similar neural command for the production of /k/ and /a/, these being supposed to share the distinctive feature of 'compactness'. Fant suggests that vowels and consonants are perceived over two separate channels. Separate subjective-distance tables for vowels and consonants were postulated by Moore (1977).

It will be seen from the above discussion of the differences in the perception of vowels and consonants that the decoding of vocalic segments begins early in the auditory block whilst their complete identification takes place either in the phonetic block[2] or at some earlier, intermediate stage.

Ades (1979) maintains that the perceptual differences between the speech signal and other auditory stimuli, just as those between vowels and consonants, are not the result of any inherent properties of these acoustic events, but rather, that they depend on a complex of stimuli accompanying them. Darwin (1976) sees no relevance - for either the cerebral specialization or for the Precategorical Auditory Store - of the vowel-consonant distinction. Dorman (1979) demonstrated that a given stimulus may be processed in the speech mode or received exactly like any other acoustic signal depending on the actual situation in

---

[1] A duration of the order of one or two larynx periods is sufficient for a vowel to be identified if it has been uttered in isolation or if the signal has been separated from the steadiest part of the vocalic segment of a nonsense word uttered slowly and carefully. When the linguistic material is continuous speech, or isolated, naturally spoken words, then even the entire vocalic segment isolated from the rest may sometimes not be enough for the vowel to be correctly decoded perceptually (Fischer-Jørgensen 1973, Łobacz 1977).

[2] The 'phonetic block' is here meant to include all aspects of processing at the phone level, the level of distinctive features as well as that of the phonemes.

which perception takes place.

2.1.6. Categorical perception.

The concept of categorical perception was introduced by the founders of the motor theory. Perception of at least some of the speech sounds in terms of special categories is one of the strongest arguments for the assumption of separate, specialized processing in the speech mode. Categorical perception implies a b s o l u t e (essentially context-independent) identification of stimuli: "Phoneme perception must be absolute or very nearly so if language is to be phonemic" (Liberman et al. 1963, p.D3). The main thesis of a number of studies in this area is that the listener can correctly distinguish speech sounds only to the extent that he identifies them as different phonemes. Discrimination is accordingly also absolute in speech since the range of variation of the speech signal has to be divided into a finite number of discrete entities (Flanagan 1972). Liberman et al. (1957) worked out a method of determining, on the basis of results of identification tests, predicted discrimination functions and of comparing them with functions actually obtained in an experiment. Signals were synthesized containing acoustic cues for a continuum /b→d→g/. The parameters of the signal related to those acoustic cues are continuously variable, so they can be quantized in arbitrary steps for experimental purposes. In Liberman et al. (1957) 14 stimuli were synthesized along the /b→d→g/ continuum. Discrimination of the stimuli was performed using the ABX method, which consists in presenting the stimuli in triplets, where X is identical either with A or with B. Stimuli A and B differed within the triplets by one, two or three steps. For the 14 different stimuli, 72 ABX triplets were constructed. The results, in the form of empirical curves (discrimination as a function of the parameter value and the magnitude of the difference between A and B) were in sufficient agreement with the predicted curves for the authors to be able to declare discrimination to be categorical. In brief, the authors' reasoning is as follows: Articulation is discretely variable whilst the acoustic cues are continuously variable. The discrimination curves have maxima at perceptual boundaries between phonemes. Perception is discrete or c a t e g o r -

i c a l in the sense that discrimination is poor between the phoneme boundaries along acoustic parameters. There are no intermediate articulations between, say, /b/ and /d/, or between /d/ and /g/, though there are intermediate a c o u s t i c stimuli. The peaks in the quasi-continuous discrimination functions indicate that perception is dependent on articulation. In Liberman et al. (1963) the discrimination and identification tests were extended to another continuum of parameter variation, this time along the opposition /slɪt : splɪt/.Now the continuously variable factor which was subject to quantization was the duration of the silent gap following the consonant /s/. Also, listeners' ability to imitate certain stimuli was tested to see whether imitation was continuous or categorical. A combination of acoustic and electromyographic measurements showed that the responses were categorical. Discrimination tests also indicated that perception is categorical along the continuously variable parameter of VOT (Voice Onset Time) in initial /p/ : /b/-like stimuli and intervocalically in the words 'rapid' : 'rabid' (Liberman et al. 1967).

Fry et al. (1962) performed experiments on the identification and discrimination of American-English vowels. Along the /ɪ→ɛ→æ/ continuum, 13 stimuli were synthesized. For the perception of vowels, with testing methods similar to those used for consonants (cf. above p. 35) less distinct results were obtained: The listeners discriminated many intraphonemic stimuli, which lead the authors to the conclusion that the perception of vowels was not categorical. According to Liberman et al. (1964) sounds that are perceived categorically are processed faster than those perceived continuously. Liberman et al. (1967) were able to demonstrate that there is a distinction between auditory stimuli requiring special coding in speech production and reconstruction (in speech perception) and stimuli processed without either of these activities. The reconstructed elements are identified in an absolute manner in the left (the dominant) hemisphere, the others - including vowels - not being perceived categorically, need not be subject to processing in the specialized hemisphere.

In a series of articles, Lane (1965, 1968a, 1968b, 1970), in his attempts to defeat most of the tenets of the founders of the

motor theory (cf. above p. 35), including the assertions on the categorical perception of consonants, questioned the design and the interpretation of the experiments (the selection of subjects, improper averaging of results over all listeners, etc.) as well as their general conception. Performing his own experiments, Lane tried to show that identification curves for consonants are not significantly different from those for vowels or from those for non-linguistic acoustic signals. Although consonants are produced by articulations that are not continuously variable whilst vowels are articulated by continuously variable configurations of the vocal tract, yet the maxima of the discrimination curves at phoneme boundaries appear for both vowels and consonants as well as in experiments with non-linguistic stimuli. This calls in question the claim that there is a special perceptual mechanism for speech at all.

In response to Lane's criticism, the proponents of the motor theory formulate their main theses more moderately (Studdert-Kennedy et al. 1970) and submit, among other things, that the difference between vowels and consonants consists in the d e g r e e to which these are perceived either continuously or categorically. But they adhere to their proposition of a special speech processing mode. "Is phonetic perception to be explained by the principles of auditory psychophysics and discrimination learning? Lane would answer, Yes'. We say, No'". (p.248). In their view, categorical perception is the result of the working of a special decoder available to man as part of his species-bound linguistic ability.

Fujisaki and Kawashima[1] developed a formal, functional model of the processes involved in speech sound discrimination. They assumed the presence of two separate memories: auditory and phonetic, and two stages of processing depending on the concrete circumstances of speech reception. If the listener hears two phonetically different sounds, i.e., if A and B (in an ABX triplet) belong to distinctly different categories, an initial

[1] The authors' views are here reconstructed from various other sources such as Pisoni (1973, 1975) and (1971a 1978), Samuel (1970), Lehiste (1972), Darwin (1976), Studdert-Kennedy (1976a), Ainsworth (1976).

decision is made on the basis of phonetic criteria. If, however, two sounds that are subject to discrimination belong to the same phonetic type e.g., allophones of consonant phoneme or a vowel phoneme , the initial decision is made by reference to stored (purely) auditory representations of the acoustic parameters of such sounds. The final discrimination result is the effect of judgments at both levels. Fujisaki and Kawashima also demonstrated that short vowels, like vowels embedded in a constant contextual frame, are perceived more categorically than long and isolated vowels (see also Stevens 1968).

In the seventies, the question of categorical speech perception was taken up again by Pisoni in a series of publications, especially in his monograph "On the Nature Categorical Perception of Speech Sounds" (1971 a). The main tasks here are:

(1) to verify the replicability of the results obtained by the proponents of the motor theory,

(2) to study the effect of experience on the degree to which the response is categorical (testing Lane's arguments),

(3) to test Fujisaki and Kawashima's model and

(4) to find whether the form of the discrimination functions is affected by the formulation of the instructions given to the listeners.

At two points Pisoni's results differ from the findings of Liberman and his associates. First, identification of the vowels is found to be better than that of consonants. And secondly, the differences between the predicted discrimination functions and those calculated on the basis of the experimental data are greater for vowels than for consonants. The effect of learning on the discriminant processes is detectable, but weak. By performing some additional experiments, Pisoni was able to confirm Fujisaki and Kawashima's views on the role of auditory memory in the discrimination process and on the effect of stimulus duration on the extent to which the judgments were categorical - provided the traditional ABX combination is used as the selected method. He also found that the instructions affect the results (cf. also Perry and Pisoni, 1977). In Pisoni (1973) a classification of the speech perception mechanisms was proposed into

categorical, or phonetic, processing and continuous , or auditory processing. The distinction is here related to the retention time, auditory memory being a more transitory store.

Cutting (1975) tried to find out whether discrimination of nonlinguistic sounds may be acquired by practice. The sounds of selected musical instruments were the object of his experiments. He demonstrated that discrimination of musical sounds could be made as strong as the discrimination of speech sounds. A comparison of the discriminability of musical intervals, noises and buzzes as well as such phonetic parameters as Voice Onset Time was performed by Carney and his associates (Carney et al. 1977). They came to the conclusion that results obtained for the various tests neither confirmed nor refuted the thesis on the categorical nature of perception. According to Pisoni and Lazarus (1974) consonants differing by the voiced-voiceless feature can be processed on the auditory just as well as on the phonetic level, and the extent to which perception is categorical depends on the range of acoustic cues used by the listener. The greater the acoustic differences between the stimuli, the easier the identification (Pisoni and Tash 1973). The distinction between categorical and continuous perception is again found to be indistinct: "... speech stimuli which have previously been shown to be perceived in a nearly categorical mode can also be perceived in a more nearly continuous mode with relatively simple manipulations of the experimental procedure" (Carney et al. 1977, p. 333). Categorical perception turns out to apply not only to speech sounds. For any type of stimuli belonging to a class of phenomena which - on the basis of his acquired knowledge - the perceiver is able to classify, it is possible to design an experiment such that its results may be interpreted in categorical terms.

2.1.7. The feature-detector theory.

The feature-detector theory is directly related to attempts to find invariants in the acoustical signal and to define phonological distinctive feature in acoustical terms. Efforts directed towards extracting and describing invariant properties of phonetic segments belong to the class of studies that are based on the assumption that the basic unit of perception is either

the distinctive feature, or the phonetic segment or the phoneme. Stevens, in his article 'The potential role of property detector in the perception of consonants' (1975) defined the invariant acoustical properties for the phonetic feature of place of articulation. The context-independent acoustic cues, he maintains, are rapid spectral changes immediately following the release (usually the initial 20-30 ms) for the cannonical syllable CV. The invariants are relational rather than absolute. On the basis of such cues, Stevens distinguished three classes of consonants: labial, coronal and velar (plus dorsal). To determine the feature of 'place of articulation' from the primary 'context-independent' acoustical cues to which the detectors react does not require storing the features in auditory memory. The author's classification of the place of articulation into three types reflects typical configurations of the vocal tract used in the production of most consonants in most languages, so - in Stevens' opinion - two or more simple detectors are enough to distinguish many consonantal sounds. Context-dependent variables such as the frequencies and the directions of formant transitions, or the Voice Onset Time, are secondary cues learnt by the listener in the process of language acquisition. These are kept in the Precategorical Auditory Memory and made use of when the invariant cues fail. Abbs and Sussman (1971), writing on general aspects of the working of the feature detectors, state that they have a significant role to play in speech perception because they mediate in the transformation of a continuous acoustic signal into a sequence of discrete elements (phonemes, syllables, etc.).

Experimental studies of the working of the property detectors usually apply the method of selective adaptation. This method is based on the following reasoning:

If the detectors for the individual features exist, then, in a special test consisting in repeated stimulation with the same signal, they should get fatigued so that the detection treshold becomes shifted giving evidence of the detectors' adaptation. It is further assumed that the final percept, i.e., the result of analysis of the given feature, is a function of two opposite-feature detectors. The result of fatiguing only one

of them is that only the unaffected detector reacts to the event under analysis. The shift of the detection threshold is always in the direction of the unaffected detector.

There are two controversial issues related to the theory of feature detectors. The one refers to the level of processing (auditory vs. phonetic) at which the detectors function. The other concerns the mutual independence of the features processed by the individual detectors (or pairs of detectors). Most of the studies in the area of selective adaptation have dealt with two features: voicing (e.g., Eimas and Corbit 1973, Cooper W.E. 1974a) and place of articulation (e.g. Ades 1974, Bailey 1974, Cooper W.E. 1974b). Also, the effect of fatiguing the detectors with signals varying in fundamental frequency and duration as related to 'stress' or 'accent' have been investigated (Hall and Blumstein 1978).

Eimas and his associates (Eimas and Corbit 1973) as well as Stevens (1971) advocate the theory that the detectors are complex sensors reacting to l i n g u i s t i c - p h o n e t i c features such as voicing, labiality, etc. Bailey (1974) represents a different view maintaining that adaptation is related to independent a c o u s t i c cues. Ades (1976) assumes that the detectors of linguistic features are basically the same as those for any individual acoustic properties. There may only be differences in the l e v e l of processing. If they react directly to the invariants in the signal, the extracted properties being later processed in a complex fashion to yield phonetic entities, then this is tantamount to the detectors in fact working on the auditory level. If, on the other hand, they analyse a signal that has previously been re-coded, this means that they react to p h o n e t i c parameters. The object of Ades' studies of adaptation was to see whether the detectors are sensistive to individual acoustic cues or to their combinations. His results may be summarized as follows: Adaptation principally works for individual acoustic cues extracted from neural spectrograms, but there is also a possibility of adaptation to combinations of such cues. The outputs of the detectors may be directly mapped onto some linguistic categories (such as voicing). Place-of-articulation detectors are not directly related to

any particular linguistic feature. Detectors react to rapid amplitude rises or falls (for CV and VC combinations) and to spectral discontinuities. Like W.E. Cooper (1974a)or Summerfield and Haggard (1974), Ades takes the view that the detectors are not simple sensors, but complex receivers reacting to multidimensional acoustic invariants.

Eimas et al. (1978) tried to find whether the phonetic features are extracted from the signal independently of each other, or whether the analysing mechanisms interact. Using the method of selective adaptation the authors found that for the stimuli /ba, da, ma, na, va, za, wa, ja/ the phonetic-feature extraction proceeds interactively. The analysis of the 'place of articulation' is more dependent than the analysis of the 'manner of articulation', but this is a one-way interaction. There is both series and parallel processing. The results of the selective-adaptation tests disagree with Oden and Massaro's views (1978), who maintain that the acoustic cues are perceived independently and only later integrated when speech sounds are being identified.

2.1.8. Concluding remarks.

One of the main purposes of the above brief review of the issues of speech signal processing was to emphasize the difficulties surrounding attempts to draw a clear distinction between primary recognition and recognition in the speech mode. Although a conversion of ensembles of physical entities into a set of discriminative code elements is the basis of language-and-speech perception, the procedures of this transformation have not yet been definitely formulated. It is not known which patterns of the spatio-temoral prelinguistic analysis correspond to segments and features - i.e., the lowest-order linguistic units. Answers to this query are now beginning to be sought in studies of language and speech acquisition by infants (Chistovich 1971, Aslin and Pisoni 1977-78). The selected models of speech perception presented in 1.1. above mostly assume a distinct difference between the processes in the auditory and those in the phonetic block. Only few of them admit the possibility of mutual effects on the information-bearing elements by feedback between the blocks. Among the many interpretations of the functioning of the primary recognition systems, the

most controversial ones are those offered by Cole and Scott (1974) and by Schouten (1980). Cole and Scott suggest that in acoustic-signal analysis three processing systems are active, which are part integrated and part independent. They deal respectively with the invariant, the context-dependent and the overall-level envelope cues. These systems make it possible to perform correct identification of most segmental elements of the given language. To accept these authors' hypotheses would amount to claiming that specifically linguistic analysis is performed very early at the auditory level of perception. A diametrically opposite view is taken by Schouten who maintains that studies of categorical perception, feature-detector adaptation, duration of auditory processing, laterality of perception and interaction between the levels all indicate no specialized linguistic decoding in the auditory stage.

The discussion of differences between auditory analysis and perception in the speech mode may be concluded as follows: Speech sounds are elements of a special linguistic code, but this does not necessarily imply that all their features are different from those of other audio signals. Stop consonants and syllabic vowels are two extremes of the range of human speech sounds. The vowels are detected rather like some other non-linguistic signals whilst the stop consonants are decoded by systems that are more strongly specialized linguistically. Fujisaki (1979) is probably right when he suggests that the conventional distinction between the two levels - auditory and phonetic, or peripheral and central - is not adequate when related to the identification of phonetic segments. The transition from peripheral to central processing may be gradual. To a certain extent, segmentation and feature detection are performed at the auditory level. When the signal is processed in the speech mode, categorization is used for more detailed segmentation. Information obtained by the feature detectors is utilized to establish the properties of the segments so that they may be classified into phonetic categories (e.g., in terms of Distinctive Features). Processing in the auditory block and processing in speech mode are partially parallel.

3. Phonetic aspects of the reception of linguistic information.

3.1. Phonetic processing.

All the models of speech perception which were briefly reviewed in 1.1. above are more or less hypothetical. The hypotheses on which those models are based are derived from psychological premises and, to a greater extent, from linguistic premises.

Generally, linguistic analysis is a strict bottom-up procedure if it belongs to the structuralist-distributional type, or a strict top-down one if it represents one of the deductive--generative schools. Both approaches assume a distinct differentiation of the levels of analysis and propose a hierarchy of these levels. In a linguistic description, maximally complete and maximally unambiguous identification is performed at each level. This process is not necessarily paralleled in actual perception.

There are increasing amounts of experimental evidence to show that the individual blocks have to deal with only part of the linguistic information that is available to the auditory receptors. There are several reasons for this.

(1) Some information is lost in very early stages of the processing. The losses are partially determined by limitations of the functioning of the perceptual mechanism (such as limited storage capacity and storage time of the precategorical auditory memory), and they are also the consequence of the situation of message transmission such as insufficient attention to the message by the listener or random information losses. Experimental work indicates that the output of the individual processing blocks, including, in extreme cases, even the final decision block, contains errors, omissions and ambiguities.

(2) At the level of specialized linguistic processing, the individual fragments of the signal have different perceptual salience. There are two alternative conceptions of the processing of segmental speech elements (cf. Shields et al. 1974). One of them declares that the individual elements are ordered into concatenations. The relations between the elements are the result of this ordering and are not affected by structuring at other levels.

Each element of such a sequence is equally unpredictable. The other theory assumes that speech sounds are organized hierarchically. A sequence is restricted by the rhythmical pattern so that only some elements of the sequences are accented (made prominent in some linguistically motivated way). Accents fall on elements in the sequence which are in some sense more important. Some elements of the rhythm patterns are more predictable than others. Shields et al. show that the perceptual targets are the accented stressed syllables. Goldstein (1977) found in his experiments that perceptual significance is related to the degree of attention which is being progressively concentrated on the successive fragments of the input signal. Here, too, it is noted that the attentional factor prefers accented syllables which - when being decoded - show fewer phonetic errors. Moreover, the accented syllables are more instrumental in final decision-making on higher processing levels.

(3) Marslen-Wilson's experiments (1978) showed that, in the recognition of lexical units, the initial phonemes are clearly identified (usually the first three), after which the listener's apriori knowledge of syntax, semantics and the immediate preceding context of the word induces prediction of the word's remaining phonetic elements. "We are not assuming that the phoneme is always perceived or that speech perception is always phonemic, only that phonemes can be perceived and often must be perceived. The term 'phoneme' is used here in a linguistic sense and to denote the perceptual unit that is the nearest counterpart of the linguistic entity" (Liberman et al. 1968, p. 21, note).

According to Öhman (1975), full phonological identification is necessary in the process of learning to read and write, but in an adult's perception, cognitive-semantic operations are the most essential. It is not known to what extent those operations require phonetic processing.

The sequence of procedures from peripheral analysis through feature detection up to phonetic transcription in the form of a feature matrix is seen by Blumstein et al. (1979) as a conventional description of speech processing. These authors suggest that it is possible directly to derive lexical units from the acoustical word forms.

Cohen (1967) considers the phonemic analysis to be a kind of skeleton which enables the listener to interpret the acoustic quasi-continuum is terms of discrete segment sequences.

It was indicated in 2.1. above that part of the so-called 'phonetic' analysis of the signal is performed at early (more 'external') stages of message reception. There are also indications of a gradual, multi-stage segmentation of the acoustic signal into discrete phonetic elements. We also mentioned the view held by Massaro (1972, 1974), Studdert-Kennedy (1975, 1976b) and Savin and Bever (1970) on preliminary segmentation of the signal into syllables with a subsequent identification of phonemes.

In consideration of the various positions reviewed above in this section, the postulation of a specialized segmentation system in the human perception mechanism does not seem to be fully justified. In the phonetic analysis systems selected for discussion below, the problem of segmentation has not been unambiguously accounted for.

The phonetic-processing model proposed by Pisoni and Sawusch (1975) is presented here in Fig.12. The output of the sensory store is processed in a recognition device. This is a complex system consisting of four blocks: acoustic-phonetic analysis, phonetic-feature analysis, a phonetic buffer and a block of phonetic-feature merging. The signal, as processed in each of these stages, is stored in short-term memory. In this model, short-term memory is a form of transitory activation of the information contained in long-term memory. The authors assume that the acoustic-feature analysis block may function as a system of detectors for such features as those described by Stevens (1975). At the phonetic-feature analysis stage, the as yet ambiguous auditory features are interpreted in terms of distinctive phonetic features. This is the stage in which "the neural signals become language". The inventors of this model suggest that the decisions on the Distinctive Features and the Segments are made on the basis of information contained in the entire syllable. A similar view was expressed earlier by Bondarko (1969). The buffer functions as a device that stores the decisions on feature detection for each syllable. It is probably meant to

contain information on contextual effects in successive segments (cf. also Cutting and Pisoni 1976). The merging of features in the last block of the recognition device results in discrete phonetic segments. The output of the recognition device corresponds approximately to a distinctive-feature matrix.

A matrix form of the output of the phonetic-analysis block is also postulated in the works of Fant (1973), Studdert-Kennedy (1976a), Stevens and Halle (1967) and Chistovich et al. (1968). Bondarko et al. (1968) assume that the final effect of a phonetic analysis is a set of abstractional binary features and that only one definite auditory parameter may be ascribed to each feature. This hypothesis is in agreement with the hierarchical concept of speech processing but it differs from the bases of the Pisoni and Sawusch model.

Cole (1973) maintains that phonemes may be coded in the short-term memory by means of Distinctive Feature, though only some of the features characterizing a given phoneme are actually used in the decoding process.

The next stage of the analysis is the processing in the phonological block. In the Pisoni and Sawush model the phonological-analysis block operates in parallel with the syntactic system and the semantic system. The phonological analysis consists in interpreting the segmental and the prosodic information arriving from the Precategorical Auditory Store of the recognition device. The ultimate result of the analysis is, in addition to tentative phonetic segments, morpheme boundaries and word boundaries.

The rules operating on the input to the phonological block determine which phonetic segments function as distinctive elements in the language and to what extent the attributes describing those elements may be language-specific.

The difference between the phonetic level and the phonological level is of little significance for the native listener, who receives the signal in terms of functional, i.e., phonological categories of his language (cf. also Pisoni 1978).

Similar hypotheses form the basis of a phonetic-analysis model proposed by Allen and Haggard (1977), here shown in Fig. 13. There are two buffer stores in this model, one for acoustic features and the other for phonetic features. The buffers are

necessary because not all features can be extracted with the same speed and because the information on each feature must be held independently of the information on other features until a phonological analysis is being performed. Also according to Repp (1975), the extracted phonetic features are combined independently. The perceptual interdependencies between features appear in earlier stages of the process.

Jassem's system of linguistic-information processing (Jassem 1977) contains a block of acoustic-phonetic analysis and a set of blocks for phonemic classification. The output of the phonetic block is in the form of signs described as potential elements of some (yet undefined) linguistic code. These signs are described by acoustic-phonetic features strongly correlated with speech production. Processing at the phonemic level is language specific and may use two courses: either through allophonemic blocks or through Distinctive-Feature blocks. The output of the Distinctive-Feature subsystem is an array of symbols which corresponds to the 'DF matrix'. In the case of processing in the allophonic subsystem, apriori information (stored by the hearer) on phonotactic regularities in the given language is also utilized (cf. also Jassem 1979).

In Oden and Massaro's model (1978, cf. 1.1. here), the identification of a phoneme has three stages: (1) defining to what extent each possible feature has been stored in PAS, (2) matching the incoming signal against a prototype stored in long-term memory, (3) classification based on relative agreement with alternative prototypes. Within this approach, the classification of phonemes is a probabilistic process because it involves the estimation of the probability of the given speech sound being assigned to one or the other class on the basis of its affinity with one or the other prototype stored in long-term memory.

If therefore, in a given situation, perception involves processing at the phonetic-analysis level, it consists essentially in identifiying segments by their phonetic features.

According to Studdert-Kennedy (1976a), the conversion of auditory information into phonetic information has not been fully explained mainly because there is no agreement as to the function-

ing of feature detectors (see here 2.1.6). In many authors' view phonetic perception consists in converting a more or less continuous acoustic signal into elements representable as sequences of discrete phonetic symbols. The phonetic level is the first level of abstraction (cf. Studdert-Kennedy 1973).

At this stage, it is not the inherent properties of the acoustic signal that are employed, but rather the properties of the units derived from it. Phonetic transcription represents the interpretation of the signal by the speaker-hearer rather than the directly observable properties of the signal (Chomsky and Halle 1968). According to the Soviet authors (Bondarko et al. 1968), the phonetic judgment of speech stimuli is the result of auditory analysis which assigns a definite articulatory reaction to the sound. If a number of reactions to two reiterated stimuli are the same, then the two stimuli represent the same abstract phonetic "image". Cooper F.S. (1972) suggests that phonetic features are probably the important characteristics of speech production that are easily related to units derived from auditory analysis. A common articulatory-perceptual plane for feature definitions is also assumed by Stevens: "When the articulatory structures assume a certain set of positions, the sound output that results from articulatory meneuvers has a set of distinctive characteristics that can be discriminated in the auditory system from other acoustic outputs. This combination of articulatory and acoustic events defines a feature" (Stevens 1972 a, p. 49). Phonetic features are for Pisoni (1978) perceptual and mentally abstractional attributes. The features form the basis of phonetic classification.

3.2. Psychological phonetic units.

From the point of view of a linguistic interpretation of the speech-perception process, the main feature of this process is the listener's ability to organize the incoming signal into units functioning at various levels. Psycholinguists are often faced with the query which of the levels of linguistic analysis is psychologically real (Pisoni 1971a). Abstraction as a mental activity appears in many stages of speech perception, beginning at phonetic analysis and ending with complete comprehension of the decoded linguistic information. It is therefore possible to seek psychological realities at all the different levels. Particular interest attaches to the psychological reality of the phoneme

and the Distinctive Features in connection with the crucial problems of speech recognition by man as well as by machine.

Alphabetic writing came into existence long before the era of the contemporary linguistic theories which created the term and the concept of the phoneme. Yet, it has always been closely related to the sound elements of exactly that degree of abstraction. When the phoneme originated as a term and a concept late in the 19th century, it immediately became apparent that alphabetic systems in existence long before were based on the differentiation of this kind of functional speech sound units (Jones 1957). It is difficult to escape the conclusion that the phoneme has always been an element of linguistic consciousness or organized subconsciousness at least in languages using alphabetic writing and has therefore had a sociolinguistic existence. The author of the theoretical concept of the phoneme was J.Baudouin de Courtenay who initially related this linguistic entity to morphological analysis thus forecasting the later morphoneme. In his famous "Versuch einer Theorie phonetischer Alternationen" (Baudouin de Courtenay 1985) he clearly defined the phoneme as a mental element "existing in the soul" and reflecting the s e n s a t i o n of sound. Perceptual implications are obvious. In later decades various definitions of the phoneme were conceived. They may be classified as (1) psychological, (2) physical, (3) semantic, (4) distributional, (5) metaphysical, and (6) logical. All the more important definitions have been commented on by Krámsky (1974) and by Fischer-Jørgensen (1975). The named types of definitions are not strictly exclusive. It might be useful to attempt a strictly dichotomous classification of all the definitions into physical and psychological. Very broadly, the bases of the more detailed distinction into the 6 types named above would be as follows: The phoneme is

(a) a mental representation of certain speech sounds,
(b) an invariant in certain otherwise different acoustic events,
(c) a phonetic element capable of distinguishing meanings,
(d) a class of mutually exclusive and freely varying phones,
(e) a fiction created for convenience of linguistic description,

(f) a set of segmental features specified by means of axioms, definitions and theorems in accordance with mathematical logic.

There is also a definition of the phoneme as a synchronous bundle of Distinctive Features. It is difficult to assign this definition to any of the 6 types because its interpretation is possible in terms of any of them.

Since the inception of the phoneme, a psychological trend in its interpretation has been represented by a number of prominent linguists. Baudouin de Courtenay (see above) and Ščerba were among them. The latter's views on this issue are described by Fischer-Jørgensen as follows: "...the main object of phonetic studies must be the psychological sound images. It is necessary to ask the speaker which phonetic differences are used for communication and the method must be subjective.... We see that already here both psychological reality and communicative function of speech sounds are emphasized" (Fischer-Jørgensen 1975, p. 325).

A psychological definition of the phoneme was the basis of Sapir's phonological theory (Sapir 1925, 1933). For him, the value of a speech sound depends on its function in the "psychological pattern". Although Sapir also took note of the 'variants' (which figure prominently in the discributional definitions), as well as of the morphological functions of the phonemes, his main interest centred around the mental processes related to the perception of speech sounds. He believed that no conceptual entities in human experience are definable as the sum of their physical properties. For Sapir, the most appropriate method of finding phonemes was to register the reactions of native listeners to authentic speech stimuli in the process of their perception.

The phoneme as a mental element is used by some generative phonologists, (eg. by Schane 1971).

A physiological conception of the phoneme, with behavioural overtones, was submitted by Bever (1970): "...phonemes are neither perceptual nor articulatory entities. Rather, they are psychological entities of a nonsensory, nonmotor kind, related by complex rules to stimuli and to articulatory movements; but they are not a unique part of either system of directly observ-

able speech processes. In short, phonemes are behaviorally abstract. Just by virtue of standing neutrally between the behavioral systems of sensory input and articulatory output, they can interrelate these perceptual and expressive speech processes" (Bever 1970, p.13). In this author's view, the following factors support the psychological reality of phonological units: (a) the existence of alphabetic writing codes (cf. p. 50) above , (b) the existence of rhymes and allieterations in unwritten poetry, (c) easy explanation of many facts of linguistic evolution in terms of phonemes, (d) easy explanation of many regularities in present-day languages in terms of units of the extent of a phoneme or a distinctive feature, and (e) slips of the tongue, especially spoonerisms.

In his paper "What is it that we perceive when we perceive speech?", Öhman (1975) wonders whether the definitions of the phoneme worked out by linguists are at all directly related to the perception of speech, which consists in the reception of a signal in the form of concrete sound. There is, according to Öhman, a definite acoustical form of every word (the 'sound of word'). The phonemes are components of such 'word sounds' that have been derived by a systematic analysis of subjective distinctive features. Phonemes are configurations, or bundles, of such features.

The psychological reality of Distinctive Features was emphasized by Jakobson (1968). For Studdert-Kennedy (1980), the abstract elements of the phoneme are psychologically real, structural elements of lexical and grammatical morphemes.

The study of the psychological reality of the various linguistic entities is related to the problem of their conscious reception. Foss and Swinney (1973), as well as Darwin (1975), are of the opinion that the individual entities obtained by linguistic analysis (such as phonemes, morphemes, words, etc.) are perceived in an automatic process anticipating awareness, whilst the percipient's attention is directed towards the meaning of the message. McNeill and Lindig (1973) regard all the levels of processing below understanding (comprehension) as 'transparent', (i.e., subconscious).

All the same, it is possible to arrange the concrete conditions

of an experiment so as to direct the subject's attention to those elements of the speech signal which, under normal speech perception conditions, are processed automatically. If, in the course of such experiments, the subjects successfully perform a discrimination or an identification task, the 'objects' may be regarded as psychologically real because they can be brought to awareness from a level of subconsciousness.

The insufficiently sharp distinction between perception, identification and awareness (Foss and Swinney's terms) as well as differences of opinion as to the direction of information flow (bottom-up vs. top-down) resulted in a controversy between Savin and Bever (1970) and Warren (1970, 1976) on the one hand and MacNeill and Lindig (1973) and Rubin et al. (1976) on the other.

Measurements of reaction time in monitoring phonemes, syllables and words in connected speech brought Savin and Bever to the conclusion that it is the syllable that is real in perception whilst phonemes are abstract mental creations that are not directly involved in perception. The reaction time was longer in the recognition of phonemes than in the recognition of syllables. MacNeill and Lindig made additional measurements of RT (reaction time) in the perception of continuous speech and they showed that if the fragments of utterances are ranked conceptually as phoneme → syllable → word → phrase → clause , then, using test stimuli of a higher order, e.g., syllables, for monitoring lower-order elements, e.g., phonemes (as Savin and Bever did), longer RTs are obtained than if the monitored elements are discovered in stimuli of the same rank. Warren (1970, 1976) studied the mechanisms of 'restoration' of phonemes replaced in words and sentences by non-speech sounds. This author also performed tests on verbal transformation of the signal (the signal being of the extent of a word) presented repeatedly for a few minutes. Illusory changes perceived by the listeners often consisted in substitutions or additions of phonemes. Results of such tests induced Warren to support Savin and Bever's main theses against McNeill and Lindig. Warren's theory is that all phonemes are products deduced from larger patterns as a result of the recognition of higher-order patterns. Foss and Swinney's

main concern was the reality of phonemic processing in perception and the ordering of analyses at different levels. They point out that if each level were independent of the preceding ones, it would be necessary to assume a number of prototypes at each level that was several orders of magnitude larger than the number of those just below it. In particular, should the syllable (or, more drastically, the word) be perceived initially and the phoneme only derived mentally by segmenting the syllable, it would be necessary to assume, for any language, at a low level of linguistic processing, a number of prototypes reaching the thousands - a very uneconomical procedure. The authors emphasize, however, that complete identification at lower levels requires feedback from the higher levels. Rubin et al.'s experiments (1976) on phoneme selection in real and nonsense words demonstrated that RT for test phonemes detected in real words is shorter that RT for test phonemes in nonsense words. Consequently, it is necessary to assume that the lexical (semantic) value of the signal affects processing at the level of the phoneme. These authors are prepared to support Foss and Swinney's thesis which says that larger linguistic units are more accessible to awareness and that they become explicit before those of lower ranks. Processes of subconscious perception are distinct from those of consicious identification.

Thus, all the tests of phoneme detection in larger units briefly reviewd above go to show that identification and probably also discrimination in the phoneme block is not an independent process. Morton and Long (1976), testing the effect of conditional probability on the identification of word-initial phonemes admit an alternative hypothesis. They do not a priori reject the possibility that phonemes may be received independently of the recognition of words. However, having made experiments on detection of individual phonemes in words and tested the effect of a delaying factor on the various phonemes of a word, these authors also reject the possibility of phoneme identification being independent of word recognition.

Jeager (1980) sought to confirm the psychological reality of the phoneme by finding out whether different units of lower phonological order (allophones) are mentally treated as belong-

ing to a phoneme if they are assumed to do so by linguistic analysis. The experiment was performed on a bottom-up basis. For one of her experiments, Jaeger assumed that the psychological reality of an entity can be demonstrated via subjects' behaviour. She conditioned her subjects by slight electrical shocks to one kind of allophone and found that they react to another allophone of the same phoneme. The subjects evidently identified various allophones of /k/ with one prototype allophone, different phonetically from the test phones. They distinguished them clearly from prototype allophones representing other (similar) consonant phonemes. Jaeger concludes that the psychological reality of the phoneme consists (at least partially) in using it as an 'organizing link' in speech perception.

In general terms, the psychological reality of an entity implies the ability to distinguish it mentally from another entity. This can only be performed by conscious or subconscious employment of distinctive features, i.e., features that are common to objects classed together whilst separating different entities (classes). Thus, phoneme recognition and feature extraction are two aspects of the same psychological activity rather than two successive real-time procedures (cf. Repp 1975).

4. Speech perception at the word level.

In most of the human speech recognition models, of which typical examples were presented in 1.1, the successive blocks representing the stages of linguistic processing have been named according to traditional linguistic terms that refer do such basic elements as phoneme, morpheme, word and sentence. The various systems differ with respect to supraphonetic processing in the choice of the kind of basic analysis in the process of perception - morphological or lexical. In the systems proposed by Fant (1973), Pisoni (1978, based on Chomsky and Halle), Klatt (1979) and also in the proposals of Studdert-Kennedy (1973) and Marslen-Wilson (1975), a word analysis block was introduced. Other models, such as those submitted by Bondarko et al. (1968) or Jassem (1979), include a morphological analysis block. For Cohen (1967, 1980) the basic unit of perception is the word as a linguistic sign mediating between things, thoughts

and emotions on the path speaker → hearer. Morton (1971), in the description of his logogen system, uses the concept of the word, with the reservation that from the point of view of information storage it would be more proper to take the morpheme as a unit. In 1980, Studdert-Kennedy considers lexical and grammatical morphemes to be the higher-order elements. The criterion of psychological reality used in establishing the perceptual role of the linguistic units has been met for words as well as for morphemes. From the point of view of a linguistic analysis, the following information may be used in processing speech at levels higher than the phoneme:

(1) structural and semantic properties of the morph,

(2) the morpheme inventory,

(3) the distributional relations between the morph and the morpheme and between the phoneme and the morpheme,

(4) the distribution between lexical and grammatical morphemes (between lexemes and semes in modern terminology),

(5) the paradigmatic and syntactic properties of morphs and words,

(6) the probabilistic properties of morphemes and words,

(7) the semantic-lexical relations (e.g., synonymy, antonymy, homonymy).

In the above list, problems of the syntax were left out intentionally, but, as the relation between the morphological and the syntactical level is not one of hierarchy, the points 1 to 7 imply partial syntactic analysis. At the present moment it is difficult to define unambiguously which kind of information is accumulated in the process of language acquisition, which is consciously used in speech perception and which may be brought to awareness only in some unusual situations of communication. Most psycholinguistic tests belong to that type of unusual situation. Experiments which test the mechanisma of the functioning of memory, the subjective frequency of lexical units as well as studies of mental processes in the area of semantic-lexical relations, have all shown the psychological reality of the processes under investigation and the linguistic elements related to them.

The level of the word (or the lexical morpheme) is that stage

in the perception of speech which mediates between the acoustical input and the hearer's complete linguistic knowledge (cf. Cohen 1967). This is the stage at which the bottom-up procedure (acoustic and phonetic analysis) interacts with the top-down procedure (the semantic-syntactic factors see also Pisoni 1978-79). Emphasizing the role of the word in speech perception, Cohen (1980) states, on the basis of Fromkin's studies, that the speaker, controlling his own speech output, checks it at the level of the word or a word-like unit. The word, as the most appropriate element for the interaction between acoustic-phonetic processing and the hearer's higher-level linguistic knowledge is studied by Bever (1970) and Marslen-Wilson and Welsh (1978). Maruszewski and Nowakowska (1970) investigated the effect of the semantics of words on the perception of their sound form. According to Foss and Swinney (1973), the words intervene in the recognition of phonemes because, they maintain, when normally listening to speech, the hearer monitors the meanings of higher-order units but does not monitor phonemes. A similar distinction between the perception of some linguistic elements and comprehension was made by Massaro (1979). According to this author, exact comprehension may be possible without exact perception. The two processes are separate stages in the treatment of linguistic information.

Winitz and his associates (1972) formulated two criteria for decoding word boundaries in utterances. These criteria are related to lexical anticipation and phonological permissibility. Lexical anticipation depends on contextual, semantic-syntactical constraints, whilst phonological permissibility (phonotactic constrains) belong to the bottom-up procedure. Jakobson (1968) claims that the recognition of a word sequence requires not only direct, immediate detection but also extrapolating anticipation and retrospective action by short-term memory. Kozhevnikov and Chistovich (quoted in Stevens and House 1972) established that while a phrase is being processed, some words are identified at the same time as they are analysed in the auditory receptors, whereas some features of other words are held in the memory and the final decision is postponed. These authors found that if the stored information exceeds a sequence of seven syllables, some information is lost so that the listener has to guess such ele-

ments of the utterance.

Apart from the phonetic and semantic-syntactical factors active in the process of decoding words, a significant role is played by the degree of familiarity of the lexical elements. A hypothesis has been put forward that words are coded in the long-term store together with their frequencies and that the most extreme parts of the vocabulary (the most and the least frequent words) are encoded separately (cf. Sambor 1972). From the point of view of the mechanisms of perception of lexical elements, the familiarity of the element to the recipient is more important than the element's objective frequency in the language.

Broadbent (1967 and 1968) studied the effect of the word's probability on its perception. He attempted to find the factors which contribute to the preference of more frequent over less frequent words in perception. He claims that sensory information recalls from memory, as the best word candidates, the ones which might have been uttered by the speaker. Martin and Schultz (1973) studied the discrimination of words by subjects with impaired hearing. It was found that when the subjects received the signals incompletely, they substituted words of similar phonetic structure and, at the same time, of high occurence frequency in their language. Savin (1963) experimentally verified the hypothesis on the much greater reception accuracy of familiar words as compared with unfamiliar ones, in the same condition of masking by noise. Considering the response bias caused, to some extent, by the experiment as well as the fact that some acoustical properties of the speech signal are much more susceptible to masking than others, the author established the following regularities: (1) no effect of word frequency on the recognition of longer words, which typically had a higher threshold of comprehension, (2) incorrect responses were more familiar words than the stimulus words. Stowe et al. (1963) studied the influence of the sentential context on the recognition of monosyllabic nouns. They wished to find to what extent the preceding information reduces the number of possible choices. It was established that the average intelligibility of words masked by noise declines with the reduction of the length of the preceding context,

whilst the context reduces the number of alternative choices.

The following subsection discusses selected models of lexical processing in which all the three factors that are significant for word recognition are included, viz.phonemic analysis, semantic-syntactic constraints and occurence frequency of words.

4.1 Models of lexical access.

The term lexical access usually refers to the availability of lexical information stored in long-term memory. In the exemplary models presented below, the basic element of speech processing is the word. The individual systems differ in the manner in which the information about these elements is being collected, ordered and recalled.

The simplest form of the lexical access model is,in Forster's (1976) opinion, one that contains the words ordered according to the orthographic or phonetic alphabet. The first letter (or speech sound) is first looked up, then the second and so on. The procedure is the same as that used for most standard dictionaries. An improved or "pruned" version of this model is a system in which only phonotactically or graphotactically admissible sequences are stored. Forster considers such models to be uneconomical. They also introduce difficulties in defining word boundaries. The system actually proposed by this author is called a "search model". In it, the words are randomized. Each of them has its own detector. which reacts to characteristic features related to the general "impresion" of the word. When the detector begins to operate, the probability of several similar words being chosen is increased and then proper recognition follows. Forster's system contains three parallel direct access files: orthographic, phonological and semantic-syntactical as well as one main indirect file. In all the files, the words are ordered according to their frequency. The preparation of the code for the stimulus word is performed in a peripheral file. As soon as the similarity is found sufficient, the main file is accessed and it is there that the detailed comparison is performed between the features of the stimulus and the words stored in the main file. The internal detectors are so interconnected that when a certain word is activated,other semantically related words may become activated also. This model has

partly been tested experimentally. The results were ambiguous with respect to the effect of the frequency of words, letter or letters clusters on recall time. The author believes that most of the useful information is contained in the initial and the final letters or sounds.

Another typical system of lexical access is Morton's logogen model mentioned in chapter 1 (Morton 1971,1969, Morton and Broadbent 1967). The most essential part of the model is the system of logogens, these being devices that accept visual and acoustic information from the sensory mechanisms as well as the semantic-contextual information about the individual words or morphemes. There is some similarity between the logogens and the word detectors in Forster's model since the logogens, like the detectors, react to information contained in the stimuli. The logogens cumulate information until it exceeds their thresholds, after which a word response is released and sent to the buffer. Along with the sensory and the contextual properties of the stimulus, the operation of the logogens is affected by the word frequencies. The logogens of the words with high occurence frequency in the recipient's idiolect have a lower threshold of sensory information cumulation. Morton's model was developed formally and also partially tested in psycholinguistic experiments. The object of one such experiment was to examine the factors that lower the threshold values and the relative contribution of the sensory and semantic information to the recognition of words. In this work (Murrel and Morton 1974) the object of detailed study was the effect of learning different words containing either the same lexical morpheme or the same visual (phonetic) form of different morphemes on the accuracy of the response. It was found that training (memorization) lowers the threshold values of the logogens. Moreover, better results were obtained for same lexical morphemes with differing suffixes than for semantically unrelated words with similar visual (acoustic) form .

The search process in Forster's model is of a strictly bottom-up nature. In the logogen model there is interaction of bottom-up and top-down processes. But Morton's system is passive in the sense that it does not require the analysed elements

to be compared outside the system.

A third model of lexical processing was developed by Marslen-Wilson (1978), Marslen-Wilson and Welsh (1978) and was defined by the author as a model of active direct access to word recognition. The system is based on the following premises: The auditory receptors accept the initial two or three segmental elements of the word which activate simultaneously a whole class of word candidates called a word-initial cohort. This ensemble is defined exclusively by bottom-up information and includes all the words beginning with the given initial sound sequence. The activated elements of lexical memory are simultaneously informed about the conditions of the context which admits, at the particular place in the utterance, the particular word syntactically and/or semantically. The sentential context accelerates word recognition and makes a positive feedback loop possible. The top-down analysis reduces the sensibility of the system to disturbances of the acoustic phonetic input, increases the effectiveness of processing and constrains the detailed interpretation of the input information. Marslen-Wilson's model presents a generalized interpretation of many psycholinguistic experiments.

5. Some reflections on speech recognition by man and machine.

One of the first serious attepts at speech recognition by machine was made by a tandem consisting of a phonetician with strong interests in speech perception and an electronics engineer: D.B.Fry and P.B.Denes of the Phonetics Department University College London (Fry and Denes 1953) . Although working in very primitive conditions compared to the facilities of today's computer-based laboratories, they were able to obtain very encouraging results in recognizing individual phonemes using little more than a filter-bank and a simple measure based on the frequencies of spectral-energy peaks. Their relative success was largely due to the fact that the decisions were made to depend not only on the signal itself but also on the stored knowledge of transitional probabilities, i.e., a kind of primitive linguistic memory. In this way, they laid the foundations for a conception of Automatic Speech Recognition - ASR - which

held great promise. It soon became evident that any further progress in this area, depending as it did on the processing of large amounts of data, could only be obtained with the aid of computers. But there has been conspicuously little interaction between linguists, psychologists and computer scientists in the area of ASR. Engineers engaged in the development of ASR systems cannot, and do not, ignore linguistic descriptions of the speech signal, but actual co-operation in concrete projects has, ever since Fry and Denes, been virtually non-existent.

By the early 1960's most of the major ASR projects engaged computer techniques, and by the early 1970's practically no advanced ASR work was done that did not heavily rely on computers in most, or all, stages of the signal processing. A significant statement by one of the leading present specialists on ASR should be quoted in this connection: "Computers can currently do some analyses b e t t e r than humans, and some others l e s s adequately, and so a controversy continues between mathematical (statistical), information-theoretic, signal-processing or pattern-classyfying methods and human-oriented phonetic, linguistic, perceptual or neurological approaches" (Lea 1979 b, p.43-44). The implication of this statement, apparently, is that for the "mathematically" oriented projects the computer can be used more successfully. The dichotomy should not be taken too strictly. Very few projects are "purely mathematical" and hardly any "purely linguistic". But it is a fact that in the former-type work, the original aims of the scientists and experimenters have, on the hole, been realized more fully than in the latter.

Computers, as instruments, cannot be blamed for possible failures here because it is not the case that technically better machines have been used for the "mathematical" approaches. Rather, the situation seems to be due to the fact that mathematical models are more rigorous than linguistic models, if the latter are to reflect natural (rather than artificial) languages. Natural languages are by nature not quite (or even not very) rigorous and consequently less amenable to algorithmic representation. Also, at the level of applications in ASR, there is less diversity of description in the mathematical models.

The more "linguistically" oriented systems usually belong to

the type referred to as Speech Understanding Systems. Most of the better-known ones have been developed under the auspices of the Advanced Research Projects Agency (ARPA) of the US Department of Defense.

"It was observed that Harpy, the most successful of the ARPA SUR systems was more of a recognizer than an Understanding system, in that it used the acoustic data to select word sequences, admittedly within strict syntactic constraints, but with no (or almost no) semantics. Despite two decades of strong advocation of the use of higher-level linguistic information to constrain the recognition task, controversy still exists about whether you absolutely need syntax and semantics or whether our knowledge of them is sufficiently advanced to warrant their inclusion in recognition of continuous speech. Some noted that syntax, semantics and discourse increase recognition accuracy and are used by the successful human prototype systems...others argued that the basic pattern matching techniques, statistical analyses and mathematical methods (without elaborate syntax or pragmatics) have produced the best success in actual recognizers. Advocates of the more mathematical view criticized available semantic and syntactic models as ad hoc...there seems to be strong agreement that a primary area was the "front end" involving acoustic, phonetic, prosodic and phonological analyses, or the equivalent in acoustic pattern matching processes" (Lea 1979 c, p.563).

Doddington (1979) also admits that systems that are directed towards the "lower ends" of linguistic structure, such as phonemes or words (and tend to be more "mathematical") are on the whole more successful than those striving to attain higher-level recognition ("understanding") and he remarks, "A major reason for this lack of progress has been misplaced emphasis on "higher-level" information to aid the recognition progress (? process, P.Ł.), with a corresponding neglect of the development of an adequate acoustic/perceptual feature representation. During the coming years, acoustic phonetic processing will be elevated beyond its current status as a simple "phoneme server" to the intelligent recognition processor...a well-defined structure must be developed for establishing and evaluating multiple acoustical/

perceptual hypotheses regarding speech segments without mutual interference among competing hypotheses" (Doddington 1979,p.559).

Discussing various general models of speech understanding systems, Goodman and Reddy (1979) distinguish the following levels of representation: signal, parametric, segmental, phonetic, surface-phonemic, syllabic, lexical, phrasal and conceptual. Note that they are virtually the same levels as are assumed in most of the human recognition models discussed in the previous sections of the present paper. The following processes are considered by these authors: parametric, environmental, feature extraction, lexical, syntactic and semantic. Four types of models are discussed: hierarchical, goal-directed, heterarchical, blackboard and locus. In the hierarchical model information is processed in a strictly bottom-up fashion, from parametric processes to semantic processes. The goal-directed model, although refered to as representing classical "top-down" procedures, has two-way connections at each stage (top-down hypotheses being verified, at each stage, with bottom-up information). In the heterarchical model, there is a two-way flow of information between all the 6 processing blocks, i.e., $6 * (6 - 1) / 2 = 15$ two-way connections. In the blackboard model, a two-way flow of information converges on, and spreads from, an additional central processing block, the "blackboard", with connections only between the "blackboard" and all the other blocks. The "locus" model is the most complex. On the one hand, information flows from the "high-level knowledge sources" and processing blocks (semantic, syntactic, lexical, word juncture) in a converging fashion to the "integrated network". On the other hand, the "front-end" parametric information, through a "feature extraction" block and "signal-to-symbol transformation" reaches the "matching-using-beam-search-algorithm" block, which is a kind of central processor because it also receives information from the "integrated network". All the processing blocks, at least in the first three models, are replica of analogous blocks in various speech perception models discussed in the initial chapters (c.f.1.1. above). And yet, there is literally not a single reference in this article by Goodman and Reddy to any of the hundreds of articles and books devoted to problems of speech perception and human speech

processing. This is very largely symptomatic. Although a limited number of papers of a very general and largely speculative character have been jointly written by linguists and acousticians, there has been very little active contribution of either linguists or psychologists to concrete ASR or SUS projects after the early Fry-Denes co-operative effort. Computer scientists and engineers engaged in ASR and SUS have obviously tried to make use of linguistics and phonetics (not necessarily always with complete understanding), but the kind of psycholinguistics that has most often been referred to has been generative-transformational grammar and generative phonology, which have - not surprisingly - been of little help.

A notable exception to the rule of mutual near-indifference is Klatt (1979), who has drawn some very stimulating parallels between perception and automatic recognition. Indeed, his 1979 paper is probably the most important single contribution to the problem of common effects in the two otherwise different modes of speech signal processing (human and automatic) in recent years. He tries to integrate some of the most significant results of work on speech perception into his two models, which he is in the process of implementing. He presents the encouraging view that ASR can profit from psycholinguistics (not necessarily of the T-G kind) and hopes that ASR work may, in its turn, provide insights into, or suggest fruitful hypotheses on speech perception. Klatt's two models are named LAFS (Lexical Access From Spectra) and SCRIBER (cf. above here, p. 6). LAFS consists of a psychophysical filter bank and a decoding network, in series. The speech wave is the input to the former. The output of the decoding network is replaced in SCRIBER by a phonetic analyser outputting phonetic transcription. As a front-end model of speech perception, LAFS is in conflict with most of the results of psycholinguistic experimental work in that it assumes no segmental-phonetic decoding, but - instead - a direct transition from short-term spectral samples to the lexicon. For some systems of ASR this model may have certain advantages (the main one being that it circumvents the still troublesome problem of automatic segmentation into acoustic-phonetic phoneme-related elements). Indeed several recognizers have been constructed, with no small

success, which tacitly or explicitly assume this model, one well-known example being VICI (Voice Input Code Identifier, Scott 1976). Kubzdela (in preparation) has developed a system which recognizes items in a small, but arbitrary vocabulary on the basis of 0-1 time-frequency matrices very strongly resembling "visible speech" spectrograms. Otherwise, by far the greatest proportion of recognizers do apply some kind of acoustic-phonetic segmentation directly related to perceptual units representing phonemes. There are very few, if any, experimental data supporting the hypothesis that LAFS in fact reflects human speech processing. The various models of speech perception discussed in the preceding sections of the present paper distinctly tend to include phoneme-level segmentation and therefore Klatt's SCRIBER may be a better representation of human speech recognition. It should also be noted that there is at least one system, developed and improved for over 20 years specifically directed towards segmental phoneme recognition, viz. Dreyfus-Graf's. Its earliest description is contained in Dreyfus-Graf (1949) and one the latest, in Dreyfus-Graf (1976).

The following general observation may be made about relations between human and mechanical speech recognition:

(1) ASR systems have been developed which recognize either low-order linguistic units, viz. phonemes or isolated words, or high-order linguistic units such as sentences or phrases. The latter more fully reflect Human Speech Recognition - HSR.

(2) HSR draws on vast amounts of knowledge stored in various memories and the variation of the speech signal in normal man-to-man communication in natural language is incomparably greater than anything that has been attacked in ASR, so that ASR need not, and in fact cannot use nearly the same amounts of apriori knowledge.

(3) It is extremely doubtful that ASR systems will be attempted in the foreseeable future that will deal with anything nearly as varied as natural language because too little is known about speech and language for workable algorithms to be designed to do the job. At the moment it does not seem possible to imagine a practical situation in which a mechanical device (computer or otherwise) may be called upon to deal with just anything

spoken in a given language - a situation essentially typical for man (apart from specialized knowledge and specialized vocabulary). As far as can be predicted on existing evidence, mechanical devices, though increasingly efficient and versatile, will be used for specified purposes (such as spoken data input, demand for information in a particular area, etc.).

(4) The extent to which studies of human speech perception and recognition may be useful to ASR projects largely depends on the kind and the variation of the speech signal that an ASR system is designed to cope with. ASR systems which attempt to recognize high-order units must consist of several processing blocks. Al least some of these blocks have to parallel human recognition processes. But there is a large measure of agreement among specialists that more acoustic-phonetic knowledge, including perceptual strategies, should be built into any system except the simplest small-vocabulary word recognizers.

(5) Single-level phoneme recognizers may be used to perform very restricted (though nonetheless very important) tasks, but any systems going beyond that deal, and will no doubt continue to deal, at some stage or another, with lexical units.

It follows from the above observations that studies of HSR at the phonetic and lexical levels (apart from their intrinsic linguistic interest) represent the area of psycholinguistics most directly relevant to ASR.

REFERENCES

ABBS, J.H., SUSSMAN, H.M.: Neurophysiological feature detectors and speech perception: a discussion of theoretical implications, JSHR 14, 23-36, 1971.

ADES, A.E.: A bilateral component in speech perception, JASA, 56, No 2, 610-616, 1974.

ADES, A.E.: Adapting the property detectors for speech perception, New approaches to language mechanisms, Wales, R.J., Walker, E., eds., North-Holland linguistic series, 30, 55-107, 1976.

ADES, A.E.: Speech and non-speech: what have we learned?, Proceedings of the Ninth International Congress of Phonetic Sciences. Copenhagen, vol. 2, 438-444, 1979.

AINSWORTH, W.A.: Mechanisms of speech recognition, Pergamon Press, 1976.

ALLEN, J., HAGGARD, M.: Perception of voicing and place features in whispered speech: A dichotic choise analysis, Perception and Psychophysics, vol. 21, 315-322, 1977.

ASLIN, R.N., PISONI, D.B.: Some developmental processes in speech perception, Research of speech perception, Progress Report No 4, Department of Psychology, Indiana University, Bloomington, 113-166, 1977-78.

BAILEY, P.J.: Perceptual adaptation for acoustical features in speech, Speech Communication Seminar, Stockholm, vol.3, 47-53, 1974.

BAUDOUIN de Courtenay, J.: Versuch einer Theorie phonologischer Alternationen, Strassburg, 1895.

BEVER, Th.G.: The influence of speech performance on linguistic structures, Advances in psycholinguistics D'Arcais, G.B.F., Levelt, W.J.M., eds. , North Holland Amsterdam, 4-30, 1970.

BLUMSTEIN, S.E., DELGUTTE, B., HALLE, M., HENKE, W.L., KEYSER, S.J., KLATT, D.H., OHDE, R.N., PAINTER, C., PERKELL, J.S., PISONI, D.W., STEVENS, K.N., ZUE, V.W.: Studies of speech production and perception, Reprint from RLE Progress Report, No 129, M.I.T. Research Labilatory of Electronics, 1979.

BONDARKO, L.W.: The Syllable Structure of Speech and Distinctive Features of Phonemes, Phonetica, 20, 1-40, 1969.

BONDARKO, L.W., ZAGORUJKO, N.G., KOZHEVNIKOW, W.A., MOŁCHANOW, A.P., CHISTOVICH, L.: Model wosprijatija rieczi czełowiekom, izd. Nauka, Sib. odd., 1968.

BROADBENT, D.E.: Perception and comminication, New York, Pergamon Press, 1958.

BROADBENT, D.E.: Word frequency effect and response bias, Psychological Review 74 /1/, 1-15, 1967.

BROADBENT, D.E.: The role of word frequency in perception, ZPSK, B. 21, H. 1-2, 173-176, 1968.

CARNEY, A.E., WIDIN, G.P., VIEMEISTER, N.F.: Noncategorical perception of stop consonants differing in VOT, Jasa, 62, 961-970, 1977.

CHISTOVICH, L.: Problems of speech perception, Form and Substance, Phonetic and Linguistic Papers Presented to Eli Fischer - Jørgensen, Hammerich, L., Jakobson, R., Zwierner, E., eds. 83-93, 1971.

CHISTOVICH, L.: Auditory processing of speech, Proceedings of the Ninth International Congress of Phonetic Sciences, vol.1, Copenhagen, 41-48, 1979.

CHISTOVICH, L., GOLUSINA, A., LUBLINSKAJA, W., MALINNIKOWA, T., Żukowa, M.: Psychological Methods in Speech Perception Research, ZPSK, B. 21, H. 1/2, 33-39, 1968.

CHOMSKY, N., HALLE, M.: The Sound Pattern of English, New York, Evanston, London, 1968.

COHEN, A.: Speech, percepts and linguistic forms, Models for the Perception of Speech and Visual form Wathen-Dunn, W., eds. M.I.T. Press, Cambridge, Massachusetts, London, 320-325, 1967.

COHEN, A.: The word as a processing unit in speech perception, Progress Report, Institute of Phonetics, University of Utrecht, vol. 5, No 1, 33-47, 1980.

COLE, R.A.: Perceiving syllables and remembering phonemes, JSHR, 16, No 1, 37-47, 1973.

COLE, R.A., SCOTT, B.: Toward a theory of speech perception, Psychological Review 81, 348-374, 1974.

COOPER, F.S.: How Is Language Conveyed by Speech?, Language by Ear and EYE, The Relationships between Speech and Reading, /Kavanagh, J.F., Mattingly, I.G., eds./, The MIT Press, 25-45, 1972.

COOPER, F.S., LIBERMAN, A.M., HARRIS, K.S., GRUBB, P.M.: Some input-output relations observed in experiments on the perception of speech, 2-nd Inter. Cong. on Cybernetics, Namur, September 3-10, 930-941, 1958.

COOPER, W.E.: Selective adaptation for acoustic cues of voicing in initial stops, Journal of Phonetics, vol. 2, No 4, 303-313, 1974 (a).

COOPER, W.E.: Adaptation of phonetic feature analyser for place of articulation, JASA 56, 617-627, 1974 (b).

CROWDER, R.G.: Visual and auditory memory, Language by Ear and Eye, The Relationship between Speech and Reading, /Kavanagh, J.F., Mattingly, J.G., eds./, MIT Press 252-257, 1972.

CROWDER, R.G.: Representation of speech sounds in precategorical acoustic storage, JEP, 14-24, 1973.

CRYSTAL, D.: Linguistics, Penguin Books, 1971.

CUTTING, J.E.: The magical number two and the natural categories of speech and music , Status Report on Speech Research, Haskins Laboratories, SR 42/43, 189-219, 1975.

CUTTING, J., PISONI, D.: An Information-Processing Approach to Speech Perception, Status Report on Speech Research, Haskins Laboratories, SR-48, 287-326, 1976.

DARWIN, C.J.: The perception of speech, Handbook of perception /Carterett, E.C., Friedman, M.P., eds./, New York, Academic Press, 175-226, 1976.

DENES, P.B.: On the motor theory of speech perception, Proceedings of the Fifth International Congress of Phonetic Sciences /Zwirner, E., Bethge, W., eds./, Basel, New York, 252-258, 1965.

DENES, P.: Motor theory of speech perception, Proceedings of the symposium on Models for Perception of Speech and Visual Form, Wathen-Dunn, W., ed. , Cambridge M.I.T. Press, 309--314, 1967.

DIVENYI, P.L.: Some psychoacoustic factors in phonetic analysis, Proceedings of the Ninth International Congress of Phonetic Sciences, Copenhagen, vol. 2, 445-542, 1979.

DODDINGTON, G.R. Whither speech recognition? in: Lea 1979 (a), 556-561, 1979.

DORMAN, M.F.: On the identification of sine-wave analogues of

CV syllables, Proceedings of the Ninth International Congress of Phonetic Sciences, Copenhagen, vol. 2, 453-460, 1979.

DREYFUS-GRAF, S.A.: Sonograph and sound mechanics, JASA 22, 732-739, 1949.

DREYFUS-GRAF, S.A.: Recognition of coded speech Phonocodes , 1976 IEEE Intern. Conf. on ASSP, Philadelphia, 198-201, 1976.

EIMAS, P.D., CORBIT, J.D.: Selective adaptation of linguistic feature detectors, Cognitive Psychology 4, 99-109, 1973.

EIMAS, P.D., Tartter, V.C., MILLER, J.L., KENTHEN, N.J.: Asymetric dependencies in processing phonetic features, Perception Psychophysics 23, 12-20, 1978.

FANT, C.G.M.: Auditory patterns of speech, in: Proceedings of the Symposium on Models for Perception of Speech and Visual Form, Wathen-Dunn, W. ed , Cambridge M.I.T. Press, 111-125, 1967.

FANT, G.: Models of speech perception, ZPSK 21, 5-8, 1968.

FANT, G.: Speech Sounds and Features, M.I.T. Press, Cambridge, Massachusetts and London, 1973.

FISCHER-JØRGENSEN, E.: Perception of German and Danish vowels with special reference to the German lax vowels /ɪ,y,ɛ/, Annual Report of the Institute of Phonetics, University of Copenhagen, vol. 7, 143-194, 1973.

FISHER-JØRGENSEN , E.: Trends in Phonological Theory, A Historical Introduction, Copenhagen, 1975.

FLANAGAN, J.L.: Speech Analysis and Perception, Second edition, Springer Verlag, Berlin, Heidelberg, New York, 1972.

FORRIN, B., MORIN, R.E.: Reaction times for items in short and long-term memory, Acta Psychologica, 30, 126-141, 1969.

FORSTER, K.J.: Accessing the mental lexicon, in: New Approaches to Language Mechanisms, /Wales, R.J., Walker, E.C.T., eds./, Amsterdam, North-Holland, 257-287, 1976.

FOSS, D.J., SWINNEY, D.A.: On the psychological reality of the phoneme: perception, identification and consciousnes, JVLVB, 12, 246-257, 1973.

FOURCIN, A.: Perceptual mechanisms at the first level of speech processing, Preceedings of the Seventh International Congress of Phonetic Sciences, Montreal, Mouton, 48-62, 1972.

FRY, D.B.: Reaction time experiments in the study of speech processing, Nouvelles Perspectives en Phonetique, Institut de Phonetique, Université Libre de Bruxelles: Conferences et Traveaux, vol. 1, 15-35, 1970.

FRY, D.B., ABRAMSON, A.S.: EIMAS, P.D., LIBERMAN, A.M.: The identification and discrimination of synthetic vowels, Language and Speech, vol. 5, 171-189, 1962.

FRY, D.B., DENES, P.B.: Mechanical sound recognition, in: Communication Theory Cherry, E.C., ed. , 426-432, London 1953.

FUJIMURA, O.: Some remarks on the Analysis - by - Synthesis as a Model of Speech Perception, ZPSK, B. 21, H. 1-2, 48-52, 1968.

FUJISAKI, H.: Some remarks on recent issues in speech perception, Proceedings of the Ninth International Congress of Phonetic Sciences, vol. 1, Copenhagen, 93-99, 1979.

GALUNOW, V.I.: Some Aspects of Speech Perception, ZPSK 21, 43-47, 1968.

GALUNOW, V.I., CZISTOWICZ, L.A.: O swiazi motornoj tieorii s obszczej problemoj rozpoznawania rieczi, Akusticzieskij Żurnał, XI, 4, 417-426, 1965.

GLUCKSBERG, S., DANKS, J.H.: Experimental psycholinguistics, An Introduction, Laurence Erlabaum Associates, Publishers, 1975.

GOODMAN, G., REDDY, R.: Alternative control structures for speech understanding systems, in Lea 1979 a, 234-246, 1979.

GRESSER, J.Y., MERCIER, G.: Automatic segmantation of speech into syllabic and phonemic units: application to French words and utterence, Auditory Analysis and Perception, Fant, G., Tatham, M.A.A., eds. , 359-382, 1975.

HAGGARD, M.: Pattern maskings and the speech perception process, Speech Communication Seminar, Stockholm, vol. 3, 39-45, August 1-3, 1974.

HAGGARD, M.: Understanding speech understanding, Structure and Process in Speech Perception, Cohen, A., Nooteboom, S.G., eds. , Springer-Verlag, 3-15, 1975.

HALL, L.L., BLUMSTEIN, S.E.: The effect of syllabic stress and syllabic organization on the identification of speech sounds, Perception and Psychophysics, 24, No 2, 137-144, 1978.

HALLE, M., STEVENS, K.N.: Speech recognition: A model and program for research, The Structure of Language Fodor, J.A., Katz, J.J., eds. , Englewood Cliffs, New Jersey, Prentice-Hall, 604-612, 1964.

HALLE, M., STEVENS, K.N.: Analysis by Synthesis., Proceedings of Seminar on Speech Comprehension and Processing, Wathen-Dunn, W., Woods, L.E., eds. , vol. 2, Paper D7, 1959.

HUGGINS, A.W.F.: Distortion of the temporal pattern of speech: Interruption and alternation, JASA 36, 1055-1064, 1964.

JAEGER, J.J.: The psychological reality of the phoneme revisited, Report of the Phonology Laboratory, No 5, Barkeley, 6-50, 1980.

JAKOBSON, R.: The role of phonic elements in speech perception, ZPSK 21, 9-20, 1968.

JAKOBSON, R., FANT, G., HALLE, M.: Preliminaries to Speech Analysis. The distinctive features and their correlates, Acoustic Laboratory, M.I.T., Report No 13, 1952 repr. 1955 .

JASSEM, W.: Założenia ogólmego modelu rozpoznawania mowy, Prace IPPT PAN, No 68, Warszawa, 1977.

JASSEM, W.: Wielostopniowy model rozpoznawania akustycznego sygnału mowy, XXVI Otwarte Seminarium z Akustyki, Wrocław - Oleśnica, 25-34, 1979.

JONES, D.: The history and meaning of the term phoneme, Suppl. to Le maitre phonetique, 1-20, 1957.

KLATT, D.H.: Review of the ARPA Speech Understanding project, JASA 62, 1345-1366, 1977.

KLATT, D.H.: Speech perception: a model of acoustic-phonetic analysis and lexical access, Journal of Phonetics, vol. 7, No 3, 279-312, 1979.

KRAMSKY, J.: The phoneme, W. Funk Verlag, München, 1974.

KRAWIEC, D., MATUSZKINA, O., MYTKOWSKI, K.: Percepcyjne granice międzysegmentalne w logatomach o budowie CVCV /C/ ze spółgłoską zwartą, Prace IPPT PAN 63, 1978.

KUBZDELA, H.: Word recognition using binary spectrograms in preparation .

KURCZ, I.: Psycholinguistyka, Przegląd problemów badawczych, PWN, Warszawa, 1976.

LANE, H.: The motor theory of speech perception, a critical review, Psychological Review, 2, 275-309, 1965.

LANE, H.L.: The hue and cry concerning categorical perception /and conversely/. ZPSK, 109-115, 1968 (a).

LANE, H.L.: On the necessity of distinguishing between speaking and listening, Proceedings of sixth International Congress of Phonetic Sciences, 1968 (b).

LANE, H.L.: Production et perception de la parole: rapports et differences, Nouvelles perspectives en phonetique, Universite Libre de Bruxelles, Conferences et Travaux, vol. 1, 87-114, 1970.

LEA, W.A. ed.: Trends in Speech Recognition, Englewood Cliffs, 1979.

LEA, W.A.: Speech recognition: past, present and future, in: Lea 1979(a), 39-98, 1979.

LEA, W.A.: Speech recognition: What is needed now? in: Lea 1979(a), 562-569, 1979.

LEHISTE, J.: The units of speech perception Ohio State University Working Papers in Linguistics, 12, 1-32, 1972.

LIBERMAN, A.M.: Some results of research on speech perception, JASA, 29, 117-123, 1961.

LIBERMAN, A.M.: Duplex perception and integration of cues: evidence that speech is different from non-speech and similar to language, Proceedings of the Ninth International Congress of Phonetic Sciences, vol. 2, Copenhagen, 467-473, 1979.

LIBERMAN A.M., COOPER, F.S., HARRIS, K.S., Mac NEILAGE, P.: A motor theory of speech perception, Proceedings of Speech Comminication, Seminar, vol. 2, Stockholm, 1963.

LIBERMAN, A.M., COOPER, F.S., HARRIS, K.S., Mac NEILAGE, P.F., STUDDERT-KENNEDY: Some observations on a model for speech perception, Proceedings of the Symposium on Models for Perception of Speech and Visual Form /Wathen-Dunn, W., ed. , 68-87, 1967.

LIBERMAN, A.M., HARRIS, K.S., HOFFMAN, H.S., GRIFFITH, B.C.: The discrimination of speech sounds within and across phoneme boundaries, JEP, 54, No 5, 358-368, 1957.

LIBERMAN, A.M., COOPER, F.S., SHANKWEILER, D.P., STUDDERT-KENNEDY, M.: Perception of the speech code, Psychological Review, 74, No 6, 431-461, 1967.

LIBERMAN, A.M., COOPER, F.S., STUDDERT-KENNEDY, M., HARRIS, K.S.,

SHANKWEILER, D.P.: On the efficiency of speech sounds, ZPSK, B. 21, HEFT 1/2, 21-32, 1968.

LICKLINDER, J.C.R.: On the process of Speech Perception, JASA 14, 590-594, 1952.

ŁOBACZ, P.: Perception of Polish vocalic segments compared to quasiautomatic recognition, Lingua Posnaniensis, XX, 97-107, 1977.

Mąc KAY, D.M.: Ways of looking at perception, Proceedings of The Symposium on Models for Perception of Speech and Visual Form, Wathen-Dunn, W., ed., Cambridge, Massachusetts, London, 25-43, 1967.

Mac KAY, D.M.: The Active-Passive' Controversery, ZPSK 21, 40-42, 1968.

MARSLEN-WILSON, W.D.: Sentence perception as an interactive parallel process, Science, vol. 189, 226-223, 1975.

MARSLEN-WILSON, W.: Linguistic Description and Psychological Assumptions in the Study of Sentence Perception, New Approaches to Language Mechanisms, Wales, R.J., WALKER, E., eds., North Holland, 203-229, 1976.

MARSLEN-WILSON, W.D.: Sequential Decisions Processes during Spoken Word Recognition, Psychonomic Society Meetings in San Antonio, Texas, November 1978, manuskrypt.

MARSLEN-WILSON, W.D., WELSH, A.: Processing interactions and lexical acces during word recognition in continuous speech, Cognitive Psychology, 10, 29-63, 1978.

MARUSZEWSKI, M.: Mowa i mózg, PWN, Warszawa, 1970.

MARUSZEWSKI, M., NOWAKOWSKA, M.: Próba eksperymentalnego badania "przezroczystości słowa dla znaczenia", Studia Psychologiczne, 10, 1970.

MASSARO, D.W.: Perceptual auditory images, JEP, 85, 411-417, 1970.

MASSARO, D.W.: Perceptual images, processing time and perceptual units in auditory perception, Psychological Review, 79, No 2, 124-145, 1972.

MASSARO, D.W.: Perceptual Units in Speech Recognition, JEP, vol. 102, No 2, 199-208, 1974.

MASSARO, D.W.: Issues in speech perception, Proceedings of the Ninth International Congress of Phonetic Sciences, Copenhagen, vol. 2, 474-481, 1979.

MASSARO, D.W., COHEN, M.M.: Perceptual auditory storage in speech recognition, Structure and Process in Speech Perception /Cohen, A., Nooteboom, S.G., eds. , Springer-Verlag, 226-245, 1975.

MATUSZKINA, O., MIKIEL, W.: Percepcyjne granice międzysegmentalne w połączeniach CVCV /C/ ze spółgłoską trącą, Prace IPPT PAN, 25/1979.

Mac NEILL, D., LINDIG, K.: The perceptual reality of phonemes, syllables, words, sentences, JVLVB, 12, 419-430, 1973.

MERMELSTEIN, P.: Automatic Segmentation of Speech into Syllabic Units, Status Report on Speech Research, Haskins Laboratories, SR-42/43. 247-256. 1975.

MILLER, G.A.: Decision units in the perception of speech, IRE Transactions on Information Theory, IT-8, 81-83, 1962.

MILLER, J.L.: Interactions in processing segmental and supra-segmental features of speech, Perception and Psychophysics, 24, No 2, 175-180, 1978.

MOORE, R.K.: Evaluating Speech Recognizers, IEEE Transaction an Acoustics, Speech and Signal Processing, ASSP-25, No 2, 178-183, 1977.

MOORE, R.K.: A multilevel system for automatic speech understanding, Speech and Hearing Work in Progress, University College, London, 91-113, 1978.

MORTON, J.: Interaction of information in word recognition, Psychological Review, 76, 165-178, 1969.

MORTON, J.: A functional model for memory, Norman, D.A. ed., Models of human memory, New York, Academic Press, 203-254, 1971.

MORTON, J., BROADBENT, D.E.: Passive versus active recognition models or is your homunculus really necessary?, Proceedings of the Symposium on Models for Perception of Speech and Visual Form, Wathen-Dunn, W., ed. , Cambridge, Massachusetts, London, 103-110, 1967.

MORTON, J., LONG, J.: Effect of word transition probability on phoneme identification, JVLVB, 15, 43-51, 1976.

MURREL, G.A., MORTON, J.: Word recognition and morphemic structure, JEP, vol. 102, No 6, 938-963, 1974.

NORMAN, D.A.: Models for human memory, Encyklopedia of linguistics, Information and Control, Meetham, A.R., Hudson, R.A., eds. , Pergamon Press, Oxford, 1969.

NORMAN, D.A.: The role of memory in the understanding of Language, Language by Ear and by Eye, The Relationship between Speech and Reading, /Kavanagh, J.F., Mattingly, J.G., eds./, M.I.T. Press, 277-288, 1972.

NORMAN, D.A.: Memory and Attention, An Introduction to Human Information Processing, 2-nd ed., New York, London, Sydney, Toronto, 1976.

NORMAN, D.A., RUMELHART, D.E.: A system for Perception and Memory, Models of Human Memory /Norman, D.A., ed./, New York, London, 21-64, 1971.

ODEN, G.C., MASSARO, D.W.: Integration of featural information in speech perception, Psychological Review, 85, 172-191, 1978.

OHMAN, S.E.G.: What is it that we perceive when we perceive speech?, Structure and Process in Speech Perception /Cohen, A., Nooteboom, S.G., eds. /, Springer-Verlag, 36-47, 1975.

PALERMO, D.S.: Development Aspects of Speech Perception: Problems for a Motor Theory, The Role of Speech in Language, /Kavanagh, J.F., Cutting, J.E., eds./, Cambridge, Massachusetts, London, 149-154, 1975.

PEREY, A.J., PISONI, D.B.: Dual Processing versus Response-limitation Accounts of Categorical Perception: A reply to Macrillan, Kaplan and Creelman, 94 Meeting of ASA, Miami Beach, 1-10, 1977.

PICKETT, J.M., POLLACK, I.: Intelligibility of experts from fluent speech: effects of rate of utterance and duration of expert, Language and Speech, vol. 6, Part 3, 151-164, 1963,

PISONI, D.B.: On the nature of categorical perception of speech sounds, Supplement to Status Report on Speech Research, Haskins Laboratories, 1971.

PISONI, D.B.: Perceptual Processing Time for Consonants and Vowels, Status Report on Speech Research, Haskins Laboratories, SR-31/32, 83-92, 1972.

PISONI, D.B.: Auditory and phonetic memory codes in the

discrimination of consonants and vowels, Perception and Psychophysics 13, 253-260, 1973.

PISONI, D.B.: Auditory short-term memory and vowel perception, Memory and Cognition, vol. 3 /1/, 7-18, 1975 a .

PISONI, D.B.: Stages of Processing in Speech Perception Eight International Congress of Phonemic Sciences, /Abstract of Papers/, 231, 1975 b .

PISONI, D.B.: Speech perception, Handbook of Learning and Cognitive Processes, vol. 6, /Estes, W.K., ed./ Hillsdale, 167-233, 1978.

PISONI, D.B.: Some measures of intelligibility and comprehension, Research on Speech Perception, Progress Report, No 5, Indiana University, 1978-79.

PISONI, D.B., LAZARUS, J.: Categorical and noncategorical models of speech perception along the voicing continuum, JASA 55, 328-333, 1974.

PISONI, D.B., SAWUSCH, J.R.: Some stages of processing in speech perception, Structure and Process in Speech Perception, Springer-Verlag, 16-35, 1975.

PISONI, D.B., TASH, J.: "Same-different" reaction times to consonants, vowels and syllables, Paper presented at the 86-th meeting of the ASA, Los Angeles, November 1, 1973.

PISONI, D.B., TASH, J.: Reaction times to comparisons within and across phonetic categories, Perception and Psychophysics, 15, 285-290, 1974.

POLLACK, I., PISONI, D.B.: On the comparison between identification and discrimination tests in speech perception, Psychonomic, Sciences, 24, 299, 1971.

REPP, B.H.: Dichotic forward and backword "masking" between CV syllables, JASA, vol. 57, No 2, 483-496, 1975.

RUBIN, P., TURVEY, M.: Van GELDER, P.: Initial phonemes are detected faster in spoken words than in spoken non-words, Perception and Psychophysics, 19, 394-398, 1976.

SAMBOR, J.: Słowa i liczby. Zagadnienia językoznawstwa statystycznego, Zakład Narod. im. Ossolińskich, 1972.

SAMUEL, A.G.: The effect of discrimination training on speech perception: noncategorical perception, Perception and Psychophysics, vol. 22, No 4, 321-330, 1977.

SAPIR, E.: The psychological reality of phonemes, Selected Writings of Edward Sapir, 46-60, 1949.

SAPIR, E.: Sound Patterns in Language, Phonology, Selected Readings, Fudge, E.C., ed. , Penguin Books, 101-114, 1973.

SAVIN, H.B.: Word-frequency effect and errors in the perception of speech, JASA 35, 200-206, 1963.

SAVIN, H.B., BEVER, T.G.: The nonperceptual reality of the phoneme, JVLVB 9, 295-302, 1970.

SCHANE, S.A.: The phoneme revisited, Language 47, 503-521, 1971.

SCHOUTEN, M.E.H.: The case against a speech mode of perception, Acta Psychologica, 44, 71-98, 1980.

SCHULTZ, M.C.: Word Familiarity Influences in Speech Discrimination, JSHR, vol. 7, No 4, 395-400, 1964.

SCOTT, P.B. VICI - A speaker-independent word recognition system, 1976 IEEE International Conf. on ASSP, Philadelphia 210-213, 1976.

SEARLE, C., JAKOBSON, J.Z., RAYMENT, S.G.: Stop consonant discrimination based on human audition, JASA 65, 799-809, 1979.

SHIELDS, J.L., Mc HUGH, A., MARTIN, J.G.: Reaction time to phoneme targets as a function of rhythmic cues in continuous speech, JEP, vol. 102, No 2, 250-255, 1974.

SHIFFRIN, R.M., ATKINSON, R.C.: Storage and retrieval processes in long-term memory, Psychological Review 76, No 2, 179-193, 1969.

SHOCKEY, L., REDDY, R.: Quantitative analysis of speech perception: Results from transcription of connected speech from unfamilar languages, Speech Communication Seminar, Stockholm, 1974.

SINGH, S.: Crosslanguage study of perceptual confusions of plosive phonemes in two conditions of distortion, JASA 40, 635-656, 1966.

STEVENS, K.N.: Towards a model for speech recognition, JASA, 32, 47-55, 1960.

STEVENS, K.N.: On the relations between speech movements and speech perception, ZPSK, 21, 102-106, 1968.

STEVENS, K.N.: Perception of phonetic segments: evidence from phonology, acoustic and psychoacoustics, The Perception of

Language /Morton, D.L., Jenkins, J.J., eds. , Columbus, 216-235, 1971.

STEVENS, K.N.: Segments, Features and Analysis by Synthesis, Language by Ear and Eye, The Relationships between Speech and Reading, Kavanagh, J.F., Mattingly, J.G., eds. , The M.I.T. Press, 47-52, 1972 a .

STEVENS, K.N.: The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data, Human Communication: A united view, David, E.E., Denes, P.B., eds. , Mac Graw-Hill Book Company, 51-66, 1972 b .

STEVENS, K.N.: The potential role of property detectors in the perception of consonants, Auditory Analysis and Perception of Speech, Fant, G., Tatham, M.A.A., eds. , Academic Press, 303-327, 1975.

STEVENS, K.N., HALLE,M.: Remarks on analysis by synthesis and distinctive features, Proceedings of Symposium on Models for Perception of Speech and Visual Form, Wathen-Dunn, W. ed. , Cambridge, M.I.T. Press, 1967.

STEVENS, K.N., HOUSE, A.S.: Speech Perception, Foundations of Modern Auditory Theory, Tobias, J., ed. , Academic, New York, vol. II, 3-62, 1972.

STEVENSON, P.W.: Reaction time measurements in speech discrimination tasks - an automated system with closed response sets, Journal of Phonetics, vol. 1, No 4, 347-368, 1973.

STOWE, A.N., HARRIS, W.P., HAMPTON, D.B.: Signal and Context Components of Word-Recognition Behaviour, JASA, vol. 35, No 5, 639-644, 1963.

STUDDERT-KENNEDY, M.: The perception of speech, Current Trends in Linguistics, vol. 12, Sebeok, T.A., ed. , Tha Hague, 2349-2385, 1973.

STUDDERT-KENNEDY, M.: From continuous signal to discrete message: Syllable to phoneme, The role of speech in Language, Kavanagh, J.F., Cutting, J.E., eds. , The M.I.T. Press, Cambridge, Massachusetts and London, 113-125, 1975.

STUDDERT-KENNEDY, M.: Speech Perception, Contemporary Issues in Experimental Phonetics, Lass, N.J., ed. , Academic Press, New York, San Francisco, London, 243-293, 1976 a .

STUDDERT-KENNEDY, M.: Universals in Linguistic Communication,

Status Report on Speech Research, Haskins Laboratories, SR-48, 43-50, 1976 b .

STUDDERT-KENNEDY, M.: Speech perception, Status Report on Speech Research, Haskins Laboratories, SR 59/60, 1-22, 1979.

STUDDERT-KENNEDY, M.: Perceiving phonetic segments, Status Report on Speech Research, Haskins Laboratories, SR-61, 123-134, 1980.

STUDDERT-KENNEDY, M., LIBERMAN, A.M., HARRIS, K.S., COOPER, F.D.: Motor theory of speech perception: A reply to Lane's critical review, Psychological Review, 77, 234-249, 1970.

STUDDERT-KENNEDY, M., SHANKWEILER, D.: Hemispheric specialization for speech perception, JASA, 48, 579-594, 1970.

SUMMERFIELD, Q., BAILEY, P.J.: What tells us that speech is speech?, Proceedings of the Ninth International Congress of Phonetic Sciences, Copenhagen, vol. 2, 482-489, 1979.

SUMMERFIELD, A.Q., HAGGARD, M.P.: Perceptual processing of multiple cues and contexts: effects of following vowel upon consonant voicing, Journal of Phonetics, vol. 2, No 4, 279-295, 1974.

SUMMERFIELD, A.Q., HAGGARD, M.P.: Vocal tract normalization as Demonstrated by Reaction Times, Auditory Analysis and Perception Fant, G., Tatham, M.A.A. eds. , 115-141, 1975.

TAPPERT, C.C.: Modeling the Processes of Speech Acquisition, ZPSK, B.21, H 1-2, 61-69, 1968.

TSEMEL, G.I.: Application in speech recognition of some data on auditory segmentation and the perception of the speech wave parameters, Auditory Analysis and Perception Fant, G., Tatham, M.A.A. eds. , 331-337, 1975.

WARREN, R.M.: Perceptual restoration of missing speech sounds, Science, 167, 392-393, 1970.

WARREN, R.M.: Auditory Illusions and Perceptual Processes, Contemporary issues in experimental phonetics Lass, N.J., ed. Academic Press, New York, San Francisco, London, 389-4177, 1976.

WICKELGREN, W.A.: Context-sensitive coding, associative memory and serial order in speech behaviour, Psychological Review, 76, 1, 1-15, 1969 b .

WINITZ, H., BELLROSE, B.: Effect of similarity of sound

substitutions on retension, JSHR 15, 677-689, 1972.
WINITZ, H., LALIVIERE, C., HERRIMAN, E.: Perception of word boundaries under conditions of lexical bias, Phonetica 27, 193-212, 1973.
WOOD, C.C.: Parallel processing of auditory and phonetic information in speech perception, Perception and Psychophysics, 15, 501-508, 1974.

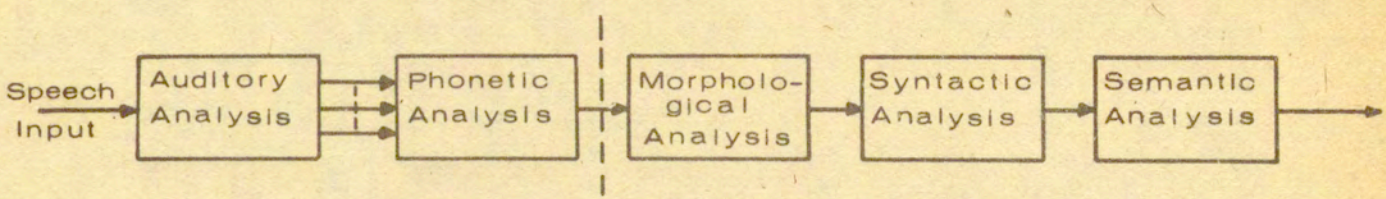

Fig. 1 A model of speech recognition by man after Bondarko et al /1966/

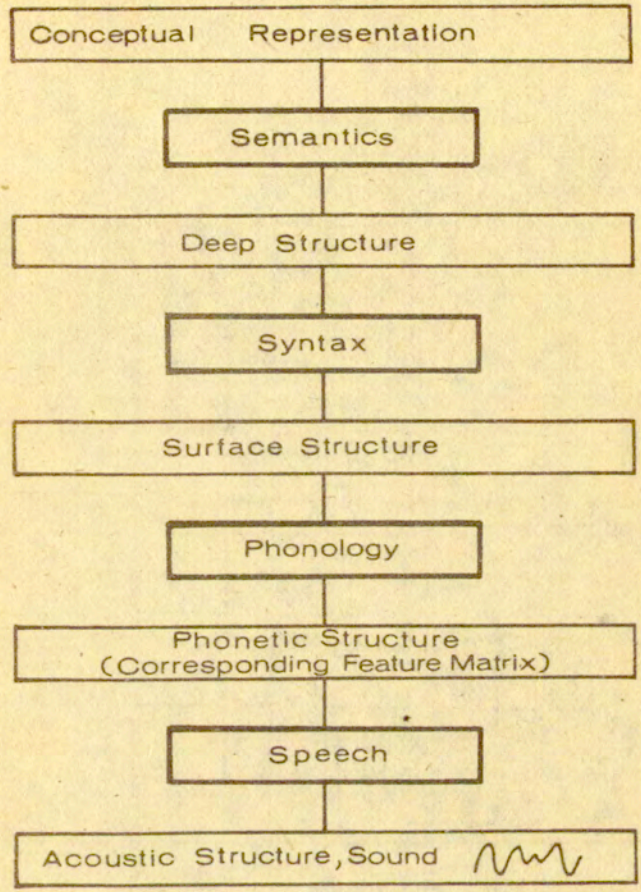

Fig. 2 Liberman's general model of linguistic-information processing /after Pisoni 1978/.

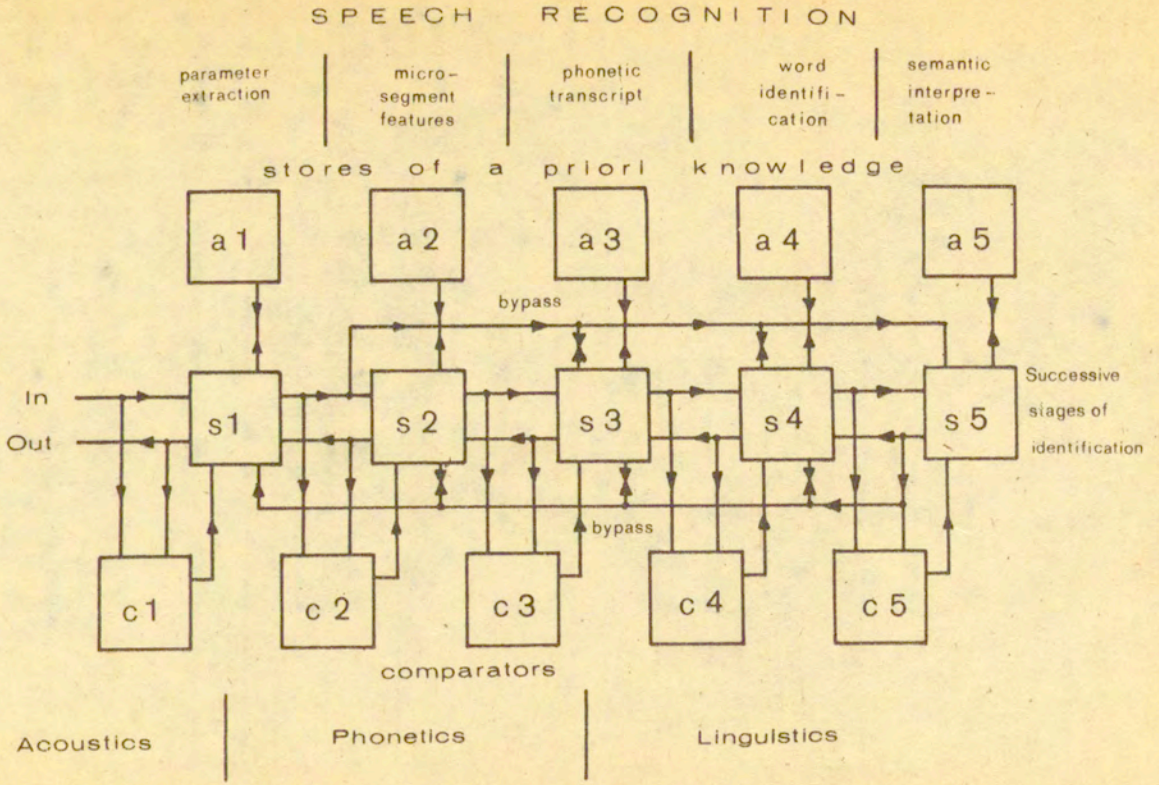

Fig. 3 Block diagram of generalized speech-recognition model after Fant /1973/.

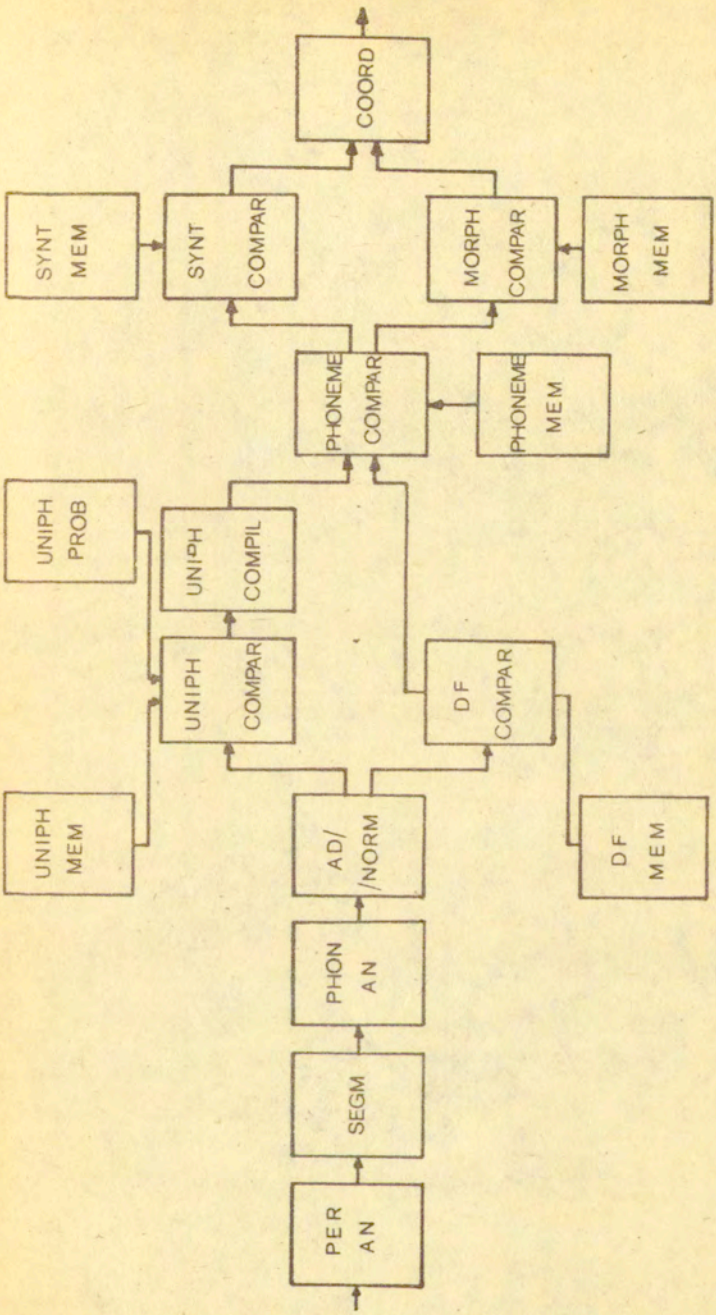

Fig.4 Jassem's /1979/ general model of speech recognition.

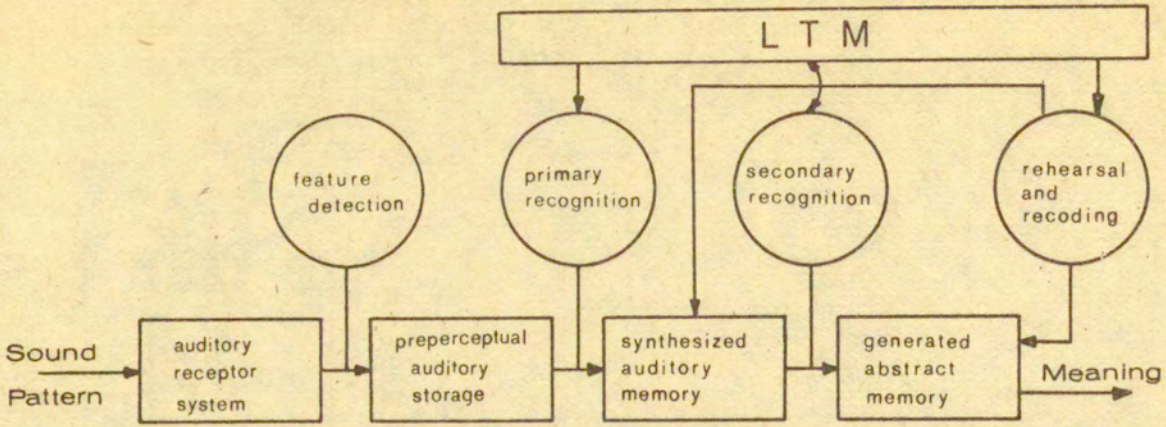

Fig. 5 Oden and Massaro's /1978/ auditory model of speech recognition.

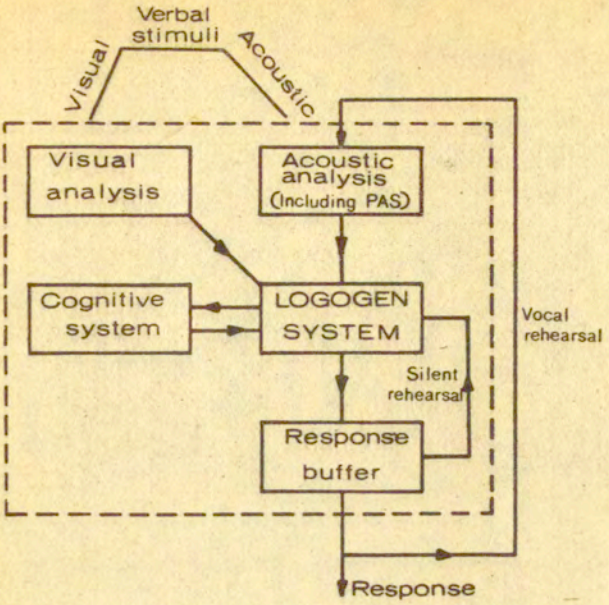

Fig.6 Morton's /1971/ logogen model of word recognition.

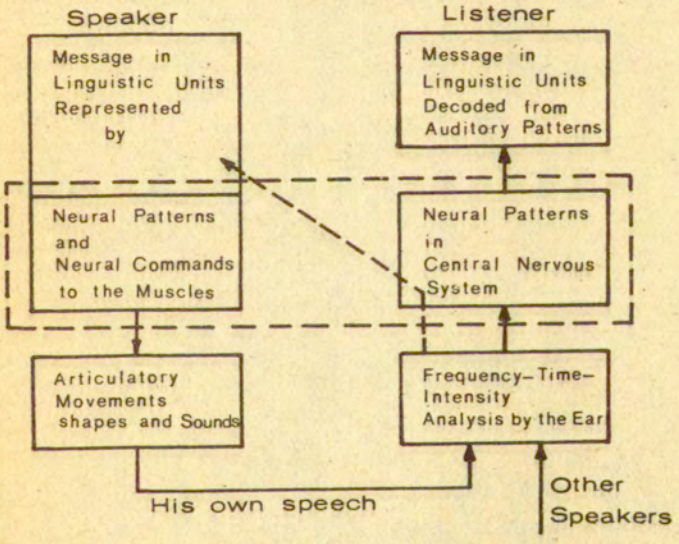

Fig.7 Libermans et al's /1968/ diagram for the speaker-hearer's procedures.

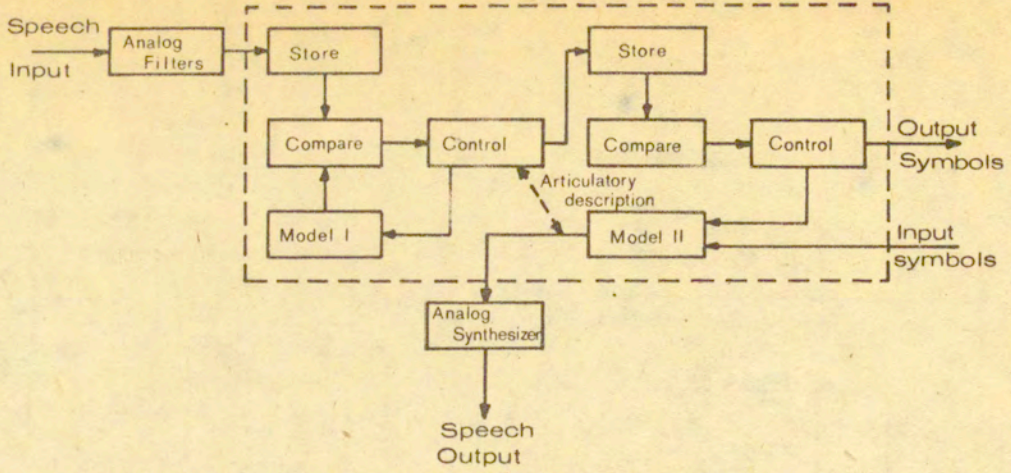

Fig.8 Steven's /1960/ information model of information processing.

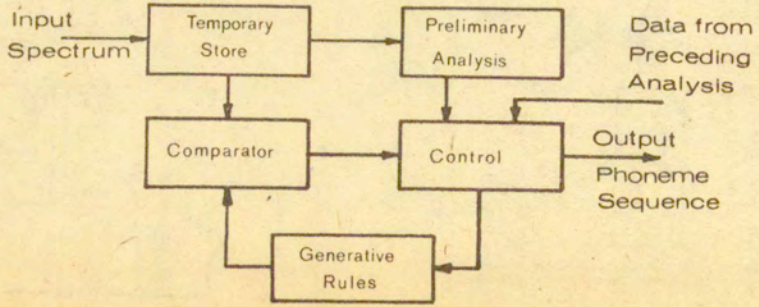

Fig.9 Halle and Stevens' /1964/ analysis-by-synthesis system.

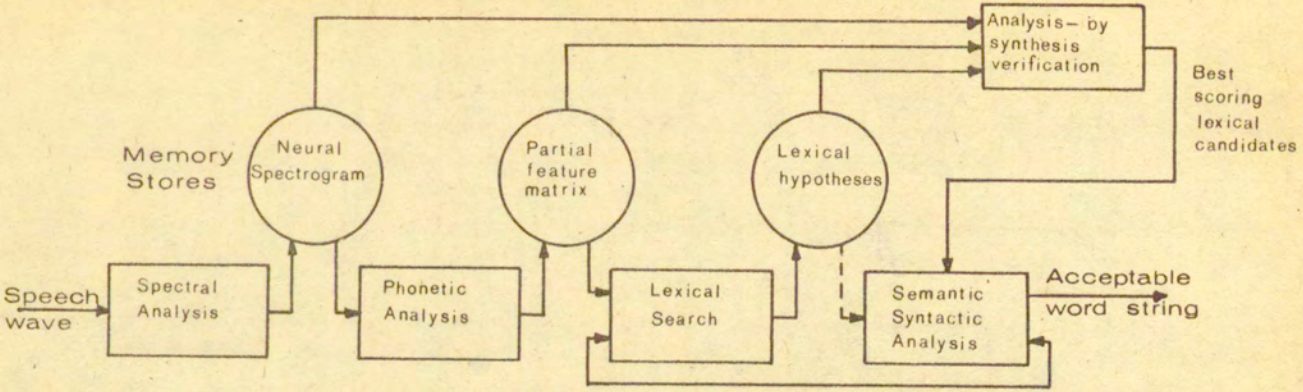

Fig. 10 Klatt's /1979/ simplified analysis-by-synthesis model.

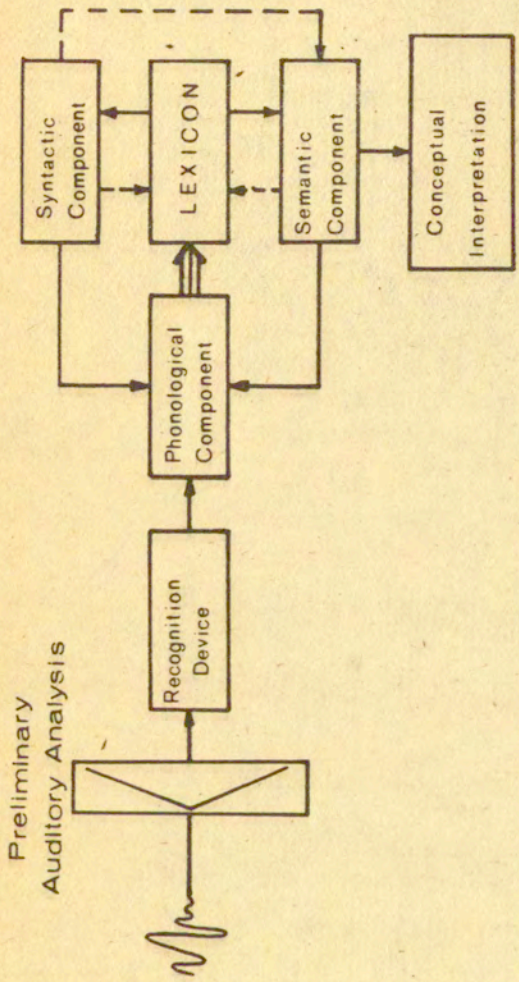

Fig.11 Chomsky and Halles' tentative model of the "novel" theory /after Pisoni 1978/.

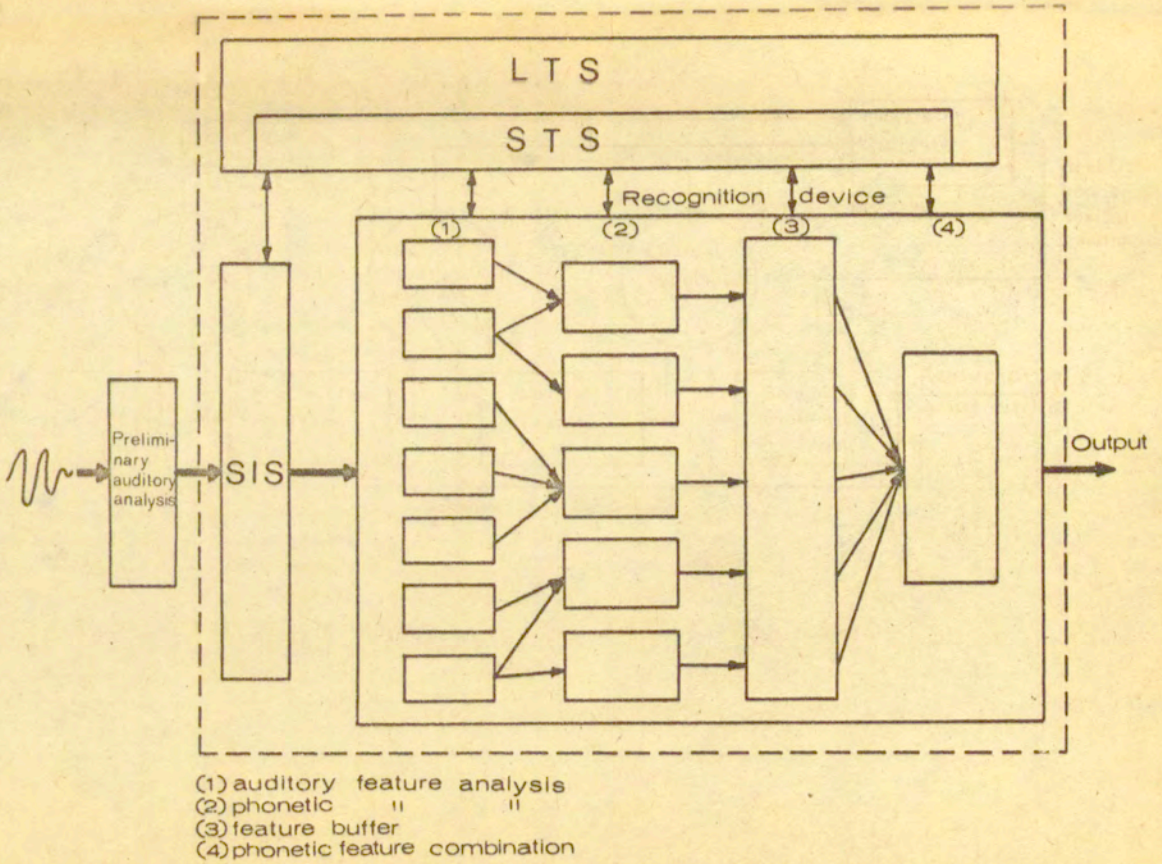

Fig. 12 The components and the organization of the recognition process after Pisoni and Sawusch /1975/.

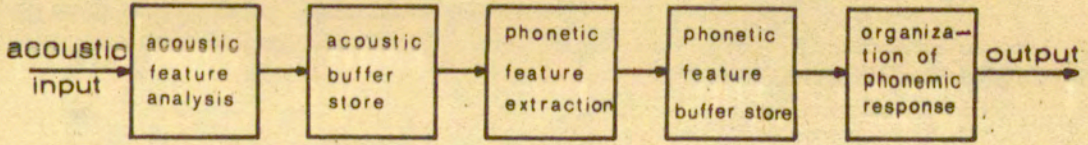

Fig. 13 A phonetic-analysis model after Allen and Haggard /1977/.