Reprinted from *Models for the Perception of Speech and Visual Form*
Weiant Wathen-Dunn, editor. 
The Massachusetts Institute of Technology. Published by The M.I.T. Press



# PASSIVE VERSUS ACTIVE RECOGNITION MODELS OR IS YOUR HOMUNCULUS REALLY NECESSARY?

A.P.U. 681/67

**John Morton and Donald E. Broadbent**
*Applied Psychology Research Unit*
*Medical Research Council*
*Cambridge, England*

MacKay (1956) has described two fundamentally different ways for automata to transform incoming "sensory" information and so recognize universals. The first of these involves a "generalising filter" and the second a "comparator." The difference between the two types may be summed up as "passive" versus "active."

The Halle-Stevens (1962) model for speech recognition involves a comparison of the input spectrum with some internally generated signals, and an error signal fed back to the generator for the next stage of the analysis-by-synthesis. The Haskins model (Liberman, 1957; Liberman, Cooper, Harris, and MacNeilage; 1962), " A motor theory of speech perception," appears to be similar in principle, if one takes the step of translating into a flow diagram statements such as "the speech sounds are perceived by reference to the articulatory movements that produce them" or "articulatory movements and their sensory feedback (or, more likely, the corresponding neurological processes) become part of the perceiving process, mediating between the acoustic stimulus and its ultimate perception."

These two active models appear to differ in one respect. In order to compare two signals meaningfully, it is essential that the signals, or neural pattern, should be at the same level. In the Halle-Stevens model the comparison is at the neuroacoustic level. A generator produces a trial phoneme sequence which is converted into a comparison spectrum presumably via the corresponding motor commands. In the Haskins model, on the other hand, it would appear that the comparison takes place on the neuroarticulatory level. Presumably then, the input spectrum would have to be converted from its neuroequivalent to a neuroarticulatory form for the comparison. As Fant (1963) has remarked,"...if the auditory analysis in the hearing process has proceeded so far as to allow the proposed articulatory matching, the decoding could proceed without an articulatory reference." He then proceeds to indicate ways in which this might be accomplished.

We have probably oversimplified and have possibly misinterpreted the Haskins statements. However, we have objections to active models of speech perception in general.

These are first, that the evidence quoted in their favor is not really inconsistent with a passive explanation, and second, that an alternative passive model is of greater generality (Morton, 1964b; Morton, 1964c). Such a model, illustrated in Figure 1, has as its central feature a "dictionary" of "units" or "logogens" which correspond to words. (The dictionary is useful for other, e.g., semantic,[1] reasons, and a "word"

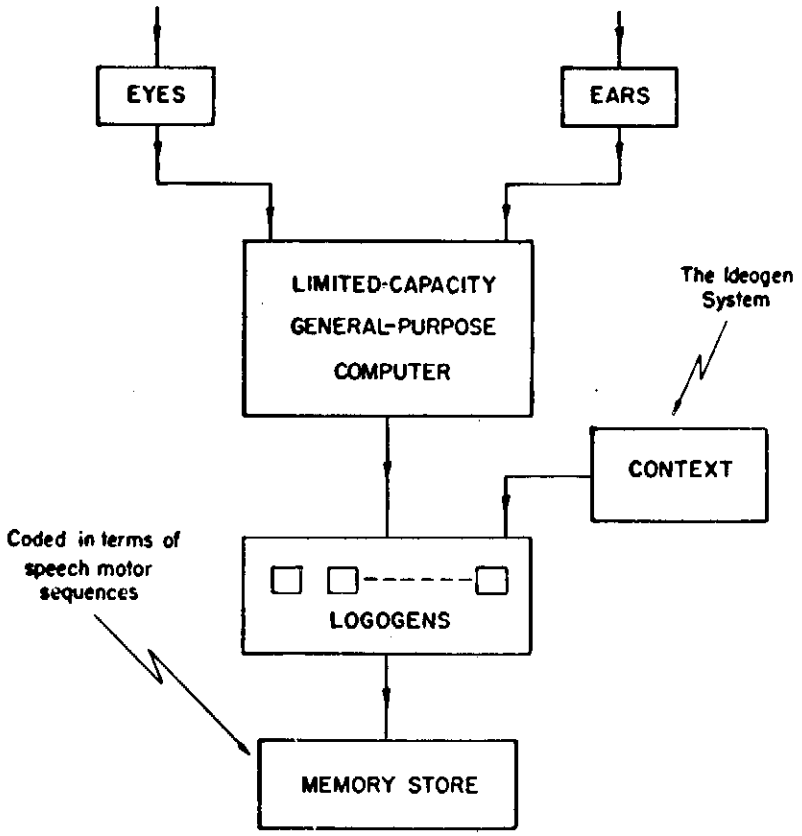


Fig. 1. Alternative passive model of speech recognition.

[1] Though it should not be thought that the logogen is the place where the meaning of a word is looked up. We have now abandoned the term "dictionary" to avoid confusion. One of us (Morton) is preparing a paper, "Rules concerning morphemes and words," which treats this more fully.

can be defined behaviorally.) When the activation in a logogen exceeds a critical level it fires, and the corresponding word is available as a response; that is, a representation of the appropriate motor sequence is stored in the Immediate Memory. This sequence would operate with or without sensory information, for the same response is available if we free associate to "chair," complete the sentence "He put the plate on the___," see the object, read the word "table" or hear it spoken. A relevant sensory input would be coded in terms of "cues," the presence of which would increase the level of activation in certain logogens. In reading, the cues would be length, shape, etc., and in listening to speech one could use the output from the Halle-Stevens Preliminary Analysis, which would utilize the Haskins acoustic-articulatory rules.

It is irrelevant to the essence of the present theory whether the information fed to the logogens is coded in terms of acoustic, articulatory or, for example, distinctive feature variables, or any combination of these, continuous or discontinuous.

In the presence of a context, which may be a word, or a sentence, or even nonverbal, certain logogens are activated differentially either by direct connections with other logogens or via higher-order nodes which will be referred to later. Since the logogens are conceived of as having the properties of signal detection units, the difference between a generation situation and a recognition situation could be merely a criterion lowering in the former case to compensate for the absence of external inputs to the system. In a recognition situation, the effect of the activation due to the context is that less sensory information is necessary to make certain logogens fire. Thus high probability words can be recognized with lower exposure times, or with a lower S/N ratio.

Such a system has the additional advantage that if I see or hear a single word in isolation the recognition procedure differs from that in context only in the amount of sensory information necessary to fire a logogen. With any active theory it would seem to be necessary either to generate all possible words (subject perhaps to limitations of length) or to have error signals fed back from the first match of such sophistication that all necessary deductions may be made as to the correct response. We feel, without a formal demonstration, that a system capable of generating such an error signal would also be capable of decoding the original neuroacoustic signal directly. In addition, the word-frequency-threshold relationship would present

difficulties for the latter of the active alternatives.[2] In a passive model this relationship is accounted for by postulating in any logogen a residual level of activation which is related to the frequency of firing of the logogen; not an outrageous postulate in a cryptoneurological theory. Miller (1964) has pointed out that the efficiency of an active model is critically dependent on the quality of the initial guess, and that if this guess is not close the listener will be unable to keep pace with rapid conversational speech. We would submit that even when such a breakdown is due to a failure of "guessing," it need not be a failure at the perceptual level (as most people understand the term) but at some higher level. The failure is to understand, not to perceive; the guess is at the idea level, not the word level. If words are missed it is equally feasible that the cause is the overoccupation with further higher-order processing of the limited capacity channel, the general-purpose computer, some short-term memory functions or attention (the terms used depending on one's taste). Everyone has surely been in the position of following every word of a speaker but understanding nothing. It has been suggested elsewhere (Morton 1964a) that the effects of redundancy on the perception of words may be traced back to higher-order units than words, where difficulties with the number of possible sequences arise as Miller, Galanter, and Pribram (1960) have remarked. Such higher-order units, called "thought units" in the previous paper, renamed "ideogens," may be few enough in number to allow the statistical properties of the language to accumulate. Connections between ideogens, and higher-order nodes can be regarded as concerned with predicting and understanding, and at such levels we have no objections to an active network for, as MacKay (1956) concluded, "...an automaton designed on statistical principles, which can evolve an internal organizing routine to respond adaptively to regularities of its sensory input, is capable in principle of developing its own symbols for concepts of any order of abstraction, including metalinguistic concepts, without prior instruction."

We suggest that the necessary statistics can be kept within reason only for the smallest units (for those conditional acoustic relationships between adjacent phonemes, so convincingly demonstrated by the Haskins work must be taken account of) and for units at a higher level

---

[2] A recent paper (Broadbent, 1966) has shown these difficulties to be major ones.

of abstraction than words; we doubt whether this can be done either for strings of phonemes (Halle and Stevens, 1962) or words.

In recognizing continuous speech, or in reading, one sees the process of prediction as a flow of information from the logogens upwards to give the context, and back down again to activate differentially the predicted words. In such a system it may be necessary to fire only the logogens corresponding to the kernel words, with some other system taking care of syntactic information in both recognition and generation. In our view the only cases in which an internal generator and comparator may operate as the sensory input would be (a) where the universals were underlearned [though Fant (1963) has questioned even this], (b) where great precision was required (as in phonetic transcription or many psychological experiments), or (c) where no response was available, e.g., with a low S/N level or high information content.

It is also worth noting that under such conditions there is a tendency to subvocalize or even vocalize not only in listening to speech but also in reading, where the input can hardly be matched actively. We cannot legitimately conclude therefore from such an observation that such activity is a part of the perceptual or recognition process, and it may equally well be a consequence of perception. Indeed, since subjectively it seems that the words most often vocalized are those with a large amount of semantic information, we may speculate as to whether subvocalization is not a form of amplification for the benefit of higher-order functions.

The passive model as presented has a place in it for other constructs which relate to other forms of behavior (Morton, 1964c). A limited-capacity channel with properties which Broadbent (1958) has described may be introduced between the primary sensory functions and the dictionary and the different properties of short- and more long-term memory (Broadbent, 1963) accounted for. In the model, material in short-term memory is coded in terms of articulation patterns. Conrad (1962) showed a correlation between errors in memory for visually presented letters and errors in hearing the letters spoken in noise. We would suggest that the former errors are due to errors in transcription of the articulatory code. Thus the alternative model covers phenomena, such as the word-frequency effect or reading, which lie outside the scope of the motor theory.

What of the evidence in favor of a motor theory? It has been criticized by Lane (1963) on methodological grounds, and he has also

shown similar results using nonspeech continua. On the present view all these results show the efficiency of speech effector coding in memory, which can apply equally to visual and auditory, speech and nonspeech stimuli, given the experience and training.

It has, for instance, often been noted that students of a foreign language have difficulty in discriminating between certain pairs of words until they have been taught to make the difference by articulatory prescription. The explanation of this phenomenon in terms of an active recognition network would be that the listener must be able to generate in order to match. We would submit that an equally valid explanation may be formulated for a passive system in that in order to discriminate between two stimuli it is necessary to have two different responses available, and until the vital articulatory distinction is available there can be no distinction between the two words. This does not of course necessarily mean that the listener cannot hear the difference between the two words spoken in juxtaposition, but here no discriminating response is required. The Haskins results seem entirely analogous to Brown and Lenneberg's findings with color perception (1954), yet a motor theory of visual color perception is not thought necessary.

To be fair we must add that the evidence quoted against a generative theory is not watertight. Roman Jakobson has already pointed out this morning that he can understand phonemic distinctions in some dialects which he cannot produce, and that children can understand before they can speak. Such observations are however not fatal for an active theory. It is possible to argue that internal discriminative responses are in a coded form which is used in the comparator system; usually this representation can be decoded to activate the speech mechanism; in certain cases, and for children, this final step cannot be taken. The comparator could still function though.

Yesterday, Attneave drew a box to represent the mechanism for form recognition with two outputs labeled "table" and "chair." He then posed the problem of how a signal on these outputs could mean something. It may mean that the appropriate logogen is activated. This, of course, does not solve the problem; it merely takes it off Attneave's desk and puts it onto ours.

Finally we might note that Halle-Stevens (1962) remark that in their system only a rough analysis may be necessary at the neuroacoustic level, with any ambiguities resolved later on the basis of constraints at the morphological, syntactic, and semantic levels. Perhaps when

we get down to details, the models will not be so far apart.

We are not vindictive; we have nothing personally against any individual homunculus, (e.g., Stevens' reference to "the active participation of the listener"); we merely believe that most of them are being given more work than is strictly necessary.

Point from the Discussion

In reply to a question from Phil Lieberman, we agree that in listening, especially to dialects, matching filters in some Primary Auditory Analysis system must be adjusted actively before we can proceed fully in the passive mode. The situation represents one of the special conditions in which we accept that a Generative System may be necessary, and it is significant that when listening to a strange dialect we often fail to perceive the first few words.

MacKay pointed out that the model is essentially a classifying system and that we know that bodily processes go on which in effect classify stimuli and indeed evoke responses without our perceiving anything. Where then, he asked, does perception live in the model? We do not think it necessary to locate "perception"; it seems sufficient to say under what conditions a subject reports "I perceived the word." In our system this would occur when a word was available as a response with a simultaneous signal to the effect that a certain amount of sensory information contributed to this event. The subject's statement may have any degree of confidence and the response need not be correct; i.e., the perception may be illusory. The difference between our model and MacKay's "raindrop model" (see his paper in this symposium) as classifiers is not that the human "perceives" but that our model can develop "its own symbols for concepts of any order of abstraction..." (see the foregoing); and this depends upon the nature of the system following the classifying system.

Al Liberman criticized the work of Lane (1963), indicating that he thought Lane's evidence to be totally inadequate. In addition, he hoped the model did not begin at the level of the word, pointing out the psychological difficulties one encounters if one enters the linguistic system at that level or even the morpheme level in reading or writing; as, for example, with logographic writing which is far more difficult than an alphabetic system. We would regard the difficulties with the logographic system to be rooted in the lack of correspondence between the two stimulus modes and between the response modes. With an

alphabetic system there is a high letter-phoneme correspondence which greatly simplifies learning, and preliminary work in England on the Initial Teaching Alphabet shows that where the spoken and written forms are in a one-to-one correspondence, the learning process is often accelerated.

References:

Broadbent, D. E. Perception and communication. London: Pergamon, 1958.

Broadbent, D. E. Flow of information within the organism. J. verb. Learn. verb. Behav., 1963, 2, 34-39.

Broadbent, D. E. Response bias and perceptual defence. Psychol. Rev., 1966, in press.

Brown, R. W., and Lenneberg, E. H. A study in language and cognition. J. abnorm. soc. Psychol., 1954, 49, 454-462.

Conrad, R. An association between memory errors and errors due to acoustic masking of speech. Nature, 1962, 193, 1314-1315.

Fant, G. Comments by G. Fant to Paper D3 A motor theory of speech perception. In Proceedings of the speech communication seminar Vol. 3. Stockholm: Royal Institute of Technology, 1963.

Halle, M., and Stevens, K. Speech recognition: A model and a program for research. IRE Trans. Info. Theory, 1962, IT-8, 155-159.

Lane, H. The motor theory of speech perception: A critical review. In Prog. Rept. No. 4, Cont. OE-3-14-013, Experimental analysis of the control of speech production and perception. Ann Arbor: Behav. Anal. Lab., Univer. Mich., 1963. (Reprint, Psychol. Rev., 1965, 72, 275-309.)

Liberman, A. M. Some results of research on speech perception. J. acoust. Soc. Amer., 1957, 29, 117-123.

Liberman, A. M., Cooper, F. S., Harris, K. S., and MacNeilage, P. F. A motor theory of speech perception. Proceedings of the speech communication seminar. Vol. 2. Stockholm: Royal Institute of Technology, 1962.

MacKay, D. M. The epistemological problem for automata. In C. E. Shannon and J. McCarthy (Eds.), Automata studies. Princeton: Princeton Univer. Press, 1956. Pp. 235-251.

Miller, G. A. The psycholinguists. Encounter, 1964, 23(1), 29-37.

Miller, G. A., Galanter, E., and Pribram, K. H. Plans and the structure of behavior. New York: Holt, 1960.

Morton, J. A model for continuous language behavior. Language and Speech, 1964, 7, 40-70. (a)

Morton, J. A preliminary functional model for language behavior. Int. Audiology, 1964, 3, 216-225. (b)

Morton, J. The effects of context on the visual duration threshold for words. Brit. J. Psychol., 1964, 55, 165-180. (c)