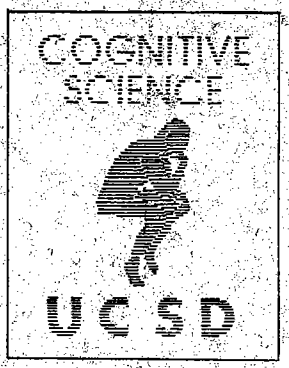
ICS REPORT NO. 8302

SPEECH PERCEPTION AS A COGNITIVE PROCESS:

THE INTERACTIVE ACTIVATION MODEL

Jeffrey L. Elman
James L. McClelland



INSTITUTE FOR COGNITIVE SCIENCE

UNIVERSITY OF CALIFORNIA, SAN DIEGO LA JOLLA, CALIFORNIA 92093

This research was conducted under Contract N00014-82-C-0374, NR 667-483 with the Personnel and Training Research Programs of the Office of Naval Research. Additional support came from grants from the National Science Foundation to Jeffrey L. Elman (BNS 79-01670) and to James L. McClelland (BNS 79-24062); an N.I.H. Career Development Award to James L. McClelland (MH 00385-02); and a grant from the Systems Development Foundation to the Institute for Cognitive Science at U.C.S.D. This support is gratefully acknowledged.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ONR-1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Speech Perception as a Cognitive Process: The Interactive Activation Model		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER ICS - 8302
7. AUTHOR(s) Jeffrey L. Elman, James L. McClelland		8. CONTRACT OR GRANT NUMBER(s) M00014-82-C-0374
9. PERFORMING ORGANIZATION NAME AND ADDRESS Center for Research in Language University of California, San Diego La Jolla, California 92093		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 667-483
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research (Code 458) Arlington, Virginia 22217		12. REPORT DATE April, 1983 13. NUMBER OF PAGES 47
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

Additional support came from grants from the National Science Foundation to Jeffrey L. Elman (BNS 79-01670) and to James C. McClelland (BNS 79-24062); an N.I.H. Career Development Award to James L. McClelland (MH00385-02); and a grant from the Systems Development Foundation to the I.C.S. at U.C.S.D.

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Interactive activation models; neural models; speech perception; speech recognition; natural language understanding; TRACE model; COHORT model.

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

OVER

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

ABSTRACT

In this paper we describe several attempts to model speech perception in terms of a processing system in which knowledge and processing is distributed over large numbers of highly interactive -- but computationally primitive -- elements, all working in parallel to jointly determine the result of the perceptual process. We begin by discussing the properties of speech which we feel demand a parallel interactive processing system, and then review previous attempts to model speech perception, both psycholinguistic and machine-based. We then present the results of a computer simulation of one version of an interactive activation model of speech, based loosely on Marslen-Wilson's COHORT model. One virtue of the model are that it is capable of word recognition and phonemic restoration without depending on preliminary segmentation of the input into phonemes. However, this version of the model has several deficiencies -- among them are excessive sensitivity to speech rate and excessive dependence on accurate information about the beginnings of words. To address some of these deficiencies, we describe an alternative called the TRACE model. In this version of the model, interactive activation processes take place within a structure which serves as a dynamic working memory. This structure permits the model to capture contextual influences in which the perception of a portion of the input stream is influenced by what follows it as well as what preceeds it in the speech signal.

Speech Perception as a Cognitive Process: The Interactive Activation Model

Jeffrey L. Elman James L. McClelland
University of California, San Diego

This report will appear in N. Lass (Ed.), *Speech and Language, Vol. 10.* Vol. 10. New York: Academic Press. Approved for public release; distribution unlimited.

This research was conducted under Contract N00014-82-C-0374, NR 667-483 with the Personnel and Training Research Programs of the Office of Naval Research. Additional support came from grants from the National Science Foundation to Jeffrey L. Elman (BNS 79-01670) and to James L. McClelland (BNS 79-24062); an N.I.H. Career Development Award to James L. McClelland (MH 00385-02); and a grant from the Systems Development Foundation to the Institute for Cognitive Sciences at U.C.S.D. This support is gratefully acknowledged. Requests for reprints should be sent to Jeffrey L. Elman, Department of Linguistics C-008; University of California, San Diego; La Jolla, California 92093.

INTRODUCTION: Interactive Activation Models

Interactive activation in speech perception
February 21, 1983
Elman & McClelland
1

Researchers who have attempted to understand higher-level mental processes have often assumed that an appropriate analogy to the organization of these processes in the human mind was the high-speed digital computer. However, it is a striking fact that computers are virtually incapable of handling the routine mental feats of perception, language comprehension, and memory retrieval which we as humans take so much for granted. This difficulty is especially apparent in the case of machine-based speech recognition systems.

Recently a new way of thinking about the kind of processing system in which these processes take place has begun to attract the attention of a number of investigators. Instead of thinking of the cognitive system as a single high-speed processor capable of arbitrarily complex sequences of operations, scientists in many branches of cognitive science are beginning to think in terms of alternative approaches. Although the details vary from model to model, these models usually assume that information processing takes place in a system containing very large numbers of highly interconnected units, each of about the order of complexity of a neuron. That is, each unit accumulates excitatory and inhibitory inputs from other units and sends such signals to others on the basis of a fairly simple (though usually non-linear) function of its inputs, and adjusts its interconnections with other units to be more or less responsive to particular inputs in the future. Such models may be called *interactive activation models* because processing takes place in them through the interaction of large numbers of units of varying degrees of activation. In such a system, a representation is a pattern of activity distributed over the units in the system and the pattern of strengths of the interconnections between the units. Processing amounts to the unfolding of such a representation in time through excitatory and inhibitory interactions and changes in the strengths of the interconnections. The interactive activation model of reading (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982) is one example of this approach; a thorough survey of recent developments in this field is available in Hinton and Anderson (1981).

In this chapter we will discuss research currently in progress in our laboratory at the University of California, San Diego. The goal of this work is to model speech perception as an interactive activation process. Research over the past several decades has made it abundantly clear that the speech signal is extremely complex and rich in detail. It is also clear from perceptual studies that human listeners appear able to deal with this complexity and to attend to the detail in

ways which are difficult to account for using traditional approaches. It is our belief that interactive activation models may provide exactly the sort of computational framework which is needed to perceive speech. While we make no claims about the neural basis for our model, we do feel that the model is far more consistent with what is known about the functional neurophysiology of the human brain than is the van Neumann machine.

The chapter is organized in the following manner. We begin by reviewing relevant facts about speech acoustics and speech perception. Our purpose is to demonstrate the nature of the problem. We then consider several previous attempts to model the perception of speech, and argue that these attempts--when they are considered in any detail--fail to account for the observed phenomena. Next we turn to our modeling efforts. We describe an early version of the model, and present the results of several studies involving a computer simulation of the model. Then, we consider shortcomings of this version of the model. Finally, we describe an alternative formulation which is currently being developed.

THE PROBLEM OF SPEECH PERCEPTION

There has been a great deal of research on the perception of speech over the past several decades. This research has succeeded in demonstrating the magnitude of the problem facing any attempt to model the process by which humans perceive speech. At the same time, important cues about the nature of the process have been revealed. In this section we review these two aspects of what has been learned about the problem.

Why Speech Perception is Difficult

* *The segmentation problem.* There has been considerable debate about what the 'units' of speech perception are. Various researchers have advanced arguments in favor of diphones (Klatt, 1980), phonemes (Pisoni, 1981), demisyllables (Fujimura & Lovins, 1978), context-sensitive allophones (Wickelgren, 1969), syllables (Studdert-Kennedy, 1976), among others, as basic units in perception. Regardless of which of these proposals one favors, it nonetheless seems clear that at various levels of processing there exist *some* kind(s) of unit which have been extracted from the speech signal. (This conclusion appears necessary if one assumes a generative capacity in speech perception.) It is therefore usually assumed that an important and appropriate task for speech analysis is somehow to segment the speech input--to draw lines separating the units.

The problem is that whatever the units of perception are, their boundaries are rarely evident in the signal (Zue & Schwartz, 1980). The information which specifies a particular phoneme is "encoded" in a stretch of speech much larger than that which we would normally say actually represents the phoneme (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). It may be impossible to say where one phoneme (or demisyllable, or word, etc.) ends and the next begins.

As a consequence, most systems begin to process an utterance by attempting what is usually an extremely errorful task. These errors give rise to further errors at later stages. A number of strategies have evolved with the sole purpose of recovering from initial mistakes in segmentation (e.g., the "segment lattice" approach adopted by BBN's **HWIM** system, Bolt, Beranek, & Newman, 1976).

We also feel that there are units of speech perception. However, it is our belief that an adequate model of speech perception will be able to accomplish the apparently paradoxical task of retrieving these units without ever explicitly segmenting the input.

Coarticulatory effects. The production of a given sound is greatly affected by the sounds which surround it. This phenomenon is termed *coarticulation.* As an example, consider the manner in which the velar stop [g] is produced in the words

gap vs. *geese*. In the latter word, the place of oral closure is moved forward along the velum in anticipation of the front vowel [i]. Similar effects have been noted for anticipatory rounding (compare the [s] in *stew* with the [s] in *steal*), for nasalization (e.g., the [a] in *can't* vs. *cat*), and for velarization (e.g., the [n] in *tank* vs. *tenth*), to name but a few. Coarticulation can also result in the addition of sounds (consider the intrusive [t] in the pronunciation of *tense* as [tɛnts].

We have already noted how coarticulation may make it difficult to locate boundaries between segments. Another problem arises as well. This high degree of context-dependence renders the acoustic correlates of speech sounds highly variable. Remarkably, listeners rarely misperceive speech in the way we might expect from this variability. Instead they seem able to adjust their perceptions to compensate for context. Thus, researchers have routinely found that listeners compensate for coarticulatory effects. A few examples of this phenomenon follow:

* There is a tendency in the production of vowels for speakers to "undershoot" the target formant frequencies for the vowel (Lindblom, 1963). Thus, the possibility arises that the same formant pattern may signal one vowel in the context of a bilabial consonant and another vowel in the context of a palatal. Listeners have been found to adjust their perceptions accordingly such that their perception correlates with an extrapolated formant target, rather than the formant values actually attained (Lindblom & Studdert-Kennedy, 1967). Oddly, it has been reported that vowels in such contexts are perceived even more accurately than vowels in isolation (Strange, Verbrugge, & Shankweiler, 1976; Verbrugge, Shankweiler, & Fowler, 1976).

* The distinction between [s] and [š] is based in part on the frequency spectrum of the frication (Harris, 1958; Strevens, 1960), such that when energy is concentrated in regions about 4kHz an [s] is heard. When there is considerable energy below this boundary, an [š] is heard. However, it is possible for the spectra of both these fricatives to be lowered due to coarticulation with a following rounded vowel. When this occurs, the perceptual boundary appears to shift. Thus, the same spectrum will be perceived as an [s] in one case, and as an [š] in the other, depending on which vowel follows (Mann & Repp, 1980). A preceding vowel has a similar though smaller effect (Hasegawa, 1976)

* Ohman (1966) has demonstrated instances of vowel coarticulation across a consonant. (That is, where the formant trajectories of the first vowel in a VCV sequence are affected by the non-adjacent second vowel, despite the intervention of a consonant.) In a series of experiments in which such stimuli were cross-spliced, Martin and Bunnell (1981) were able to show that listeners are sensitive to such distal coarticulatory effects.

* Repp and Mann (1981a, 1981b) have reported generally higher F3 and F4 onset frequencies for stops following [s] as compared with stops which follow [š]. Parallel perceptual studies revealed that listeners' perceptions varied in a way which was consistent with such coarticulatory influences.

* The identical burst of noise can cue perception of stops at different places of articulation. A noise burst centered at 1440 Hz followed by steady state formants appropriate to the vowels [i], [a], or [u] will be perceived as [p], [k], or [p], respectively (Liberman, Delattre, & Cooper, 1952). Presumably this reflects the manner in which the vocal tract resonances which give rise to the stop burst are affected during production by the following vowel (Zue, 1976).

* The formant transitions of stop consonants vary with preceding liquids ([r] and [l]) in a way which is compensated for by listeners' perceptions (Mann, 1980). Given a sound which is intermediate between [g] and [d], listeners are more likely to report hearing a [g] when it is preceded by [l] than by [r].

In the above examples, it is hard to be sure what the nature of the relation is between production and perception. Are listeners accommodating their perception to production dependencies? Or do speakers modify production to take into account pecularities of the perceptual system? Whatever the answer, both the production and the perception of speech involve complex interactions, and these interactions tend to be mirrored in the other modality.

Feature dependencies. We have just seen that the manner in which a feature or segment is interpreted frequently depends on the sounds which surround it; this is what Jakobson (1968) would have called a *syntagmatic* relation. Another factor which must be taken into consideration in analyzing features is what other features co-occur in the same segment. Features may be realized in different ways, depending on what other features are present.

If a speaker is asked to produce two vowels with equal duration, amplitude, and fundamental frequency (F0), and one has a low tongue position (such as [a]) and the other has a high tongue position (e.g., [i]) the [a] will generally be longer, louder, and have a lower F0 than the [i] (Peterson & Barney, 1952). This production dependency is mirrored by listeners' perceptual behavior. Despite physical differences in duration, amplitude, and and F0, the vowels produced in the above manner are perceived as identical with regard to these dimensions (Chuang & Wang, 1978). Another example of such an effect may be found in the relationship between the place of articulation and voicing of a stop. The perceptual threshold for voicing shifts along the VOT continuum as a function of place, mirroring a change which occurs in production.

In both these examples, the interaction is between feature and intra-segmental context, rather than between feature and trans-segmental context.

Trading relations. A single articulatory event may give rise to multiple acoustic cues. This is the case with voicing in initial stops. In articulatory terms, voicing is indicated by the magnitude of (VOT). VOT refers to the temporal offset between onset of glottal pulsing and the release of the stop. This apparently simple event has complex acoustic consequences. Among other cues, the following

provide evidence for the VOT: (1) presence or absence of first formant (F1 cut-back), (2) voiced transition duration, (3) onset frequency of F1, (4) amplitude of burst, and (5) F0 onset contour. Lisker (1957, 1978) has provided an even more extensive catalogue of cues which are available for determing the voicing of stops in intervocalic position.

In cases such as the above, where multiple cues are associated with a phonetic distinction, these cues exhibit what have been called "trading relations" (see Repp, 1981, for review). Presence of one of the cues in greater strength may compensate for absence or weakness of another cue. Such perceptual dependencies have been noted for the cues which signal place and manner of articulation in stops (Miller & Eimas, 1977; Oden & Massaro, 1978; Massaro & Oden, 1980a,b; Alfonso, 1981), voicing in fricatives (Derr & Massaro, 1980; Massaro & Cohen, 1976); the fricative/affricate distinction (Repp, Liberman, Eccardt, & Pesetsky, 1978), among many others.

As is the case with contextually governed dependencies, the net effect of trading relations is that the value of a given cue can not be known absolutely. The listener must integrate across all the cues which are available to signal a phonetic distinction; the significance of any given cue interacts with the other cues which are present.

Rate dependencies. The rate of speech normally may vary over the duration of a single utterance, as well as across utterances. The changes in rate affect the dynamics of the speech signal in a complex manner. In general, speech is compressed at higher rates of speech, but some segments (vowels, for example) are compressed relatively more than others (stops). Furthermore, the boundaries between phonetic distinctions may change as a function of rate (see Miller, 1981 for an excellent review of this literature).

One of the cues which distinguishes the stop in [ba] from the glide in [wa] is the duration of the consonantal transition. At a medium rate of speech a transition of less than approximately 50 ms. causes listeners to perceive stops. (Liberman, Delattre, Gerstman, & Cooper, 1956). Longer durations signal glides (but at very long durations the transitions indicate a vowel). The location of this boundary is affected by rate changes; it shifts to shorter values at faster rates (Minifie, Kuhl, & Stecher, 1976; Miller & Liberman, 1979).

A large number of other important distinctions are affected by the rate of speech. These include voicing (Summerfield, 1974), vowel quality (Lindblom & Studdert-Kennedy, 1967; Verbrugge & Shankweiler, 1977), fricative vs. affricate (although these findings are somewhat paradoxical, Dorman, Raphael, & Liberman, 1976).

Phonological effects. In addition to the above sources of variability in the speech signal, consider the following phenomena.

In English, voiceless stop consonants are produced with aspiration in syllable-initial position (as in [p^h]) but not when they follow an [s] (as in [sp]). In many environments, a sequence of an alveolar stop followed by a palatal glide is replaced by an alveolar palatal affricate, so that *did you* is pronounced as [dɪʃu]. Also in many dialects of American (but not British) English, voiceless alveolar stops are 'flapped' intervocalically following a stressed vowel (*pretty* being pronounced as [prɪDi]). Some phonological processes may delete segments or even entire syllables; vowels in unstressed syllables may thus be either "reduced" or deleted altogether, as in *policeman* [plismən].

The above examples illustrate *phonological processes*. These operate when certain sounds appear in specific environments. In many respects, they look like the contextually-governed and coarticulatory effects described above (and at times the distinction is in fact not clear). Phonological changes are relatively high-level. That is, they are often (although not always) under speaker control. The pronunciation of *pretty* as [prɪDi] is typical of rapid conversational speech, but if a speaker is asked to pronounce the word very slowly emphasizing the separate syllables, he or she will say [prɪ-t^hi]. Many times these processes are entirely optional; this is generally the case with deletion rules. Other phonological rules (e.g., allophonic rules) are usually obligatory. This is true of syllable-initial voiceless stop aspiration.

Phonological rules vary across languages and even across dialects and speech styles of the same language. They represent an important source of knowledge listeners have about their language. It is clear that the successful perception of speech relies heavily on phonological knowledge.

* * * * *

These are but a few of the difficulties which are presented to speech perceivers. It should be evident that the task of the listener is far from trivial. There are several points which are worth making explicit before proceeding.

First, the observations above lead us to the following generalization. There are an extremely large number of factors which converge during the production of speech. These factors interact in complex ways. Any given sound can be considered to lie at the nexus of these factors, and to reflect their interaction. The process of perception must somehow be adapted to unraveling these interactions.

Second, as variable as the speech signal is, that variability is lawful. Some models of speech perception and most speech recognition systems tend to view the speech signal as a highly degraded input with a low signal/noise ratio. This is an unfortunate conclusion. The variability is more properly regarded as the result of the parallel transmission of information. This parallel transmission provides a high degree of redundancy. The signal is accordingly complex, but--if it is

analyzed correctly--it is also extremely robust. This leads to the next conclusion.

Third, rather than searching for acoustic invariance (either through reanalysis of the signal or proliferation of context-sensitive units) we might do better to look for ways in which to take advantage of the rule-governed variability. We maintain that the difficulty which speech perception presents is not how to reconstruct an impoverished signal; it is how to cope with the tremendous amount of information which is available, but which is (to use the term proposed by Liberman et al., 1967) highly encoded. The problem is lack of a suitable computational framework.

Clues About the Nature of the Process

The facts reviewed above provide important constraints on models of speech perception. That is, any successful model will need to account of those phenomena in an explicit way. In addition, the following additional facts should be accounted for in any model of speech perception.

High-level knowledge interacts with low-level decisions. Decisions about the acoustic/phonetic identify of segments are usually considered to be low-level. Decisions about questions such as "What word am I hearing?" or "What clause does this word belong to?" or "What are the pragmatic properties of this utterance?" are thought of as high-level. In many other models of speech perception, these decisions are answered at separate stages in the process, and these stages interact minimally and often only indirectly; at best, the interactions are bottom-up. Acoustic/phonetic decisions may supply information for determining word identity, but word identification has little to do with acoustic/phonetic processing.

We know now, however, that speech perception involves extensive interactions between levels of processing, and that top-down effects are as significant as bottom-up effects.

For instance, Ganong (1980) has demonstrated that the lexical identity of a stimulus can affect the decision about whether a stop consonant is voiced or voiceless. Ganong found that, given a continuum of stimuli which ranged perceptually from *gift* to *kift*, the voiced/voiceless boundary of his subjects was displaced toward the voiced end, compared with similar decisions involving stimuli along a *giss* - *kiss* continuum. The low-level decision regarding voicing thus interacted with the high-level lexical decision.

In a similar vein, Isenberg, Walker, Ryder, & Schweickert (1980) found that the perception of a consonant as being a stop or a fricative interacted with pragmatic aspects of the sentence in which it occurred. In one of the experiments reported by Isenberg et al., subjects heard two sentence frames: *I like ____joke* and *I like ____drive*. The target slot contained a stimulus which was drawn from a

to - the continuum (actually realized as [tə] - [ðə], with successive attenuation of the amplitude of the burst + aspiration interval cueing the stop/fricative distinction). For both frames *to* as well as *the* result in grammatical sentences. However, *joke* is more often used as a noun, whereas *drive* occurs more often as a verb. Listeners tended to hear the consonant in the way which favored the pragmatically plausible interpretation of the utterance. This was reflected as a shift in the phoneme boundary toward the [t] end of the continuum for the *I like* ____ *joke* items, and toward the [ð] end for the *I like* ____*drive* items.

The role of phonological knowledge in perception has been illustrated in an experiment by Massaro and Cohen (1980). Listeners were asked to identify sounds from a [li]-[ri] continuum (where stimuli differed as to the onset frequency of F3). The syllables were placed after each of four different consonants; some of the resulting sequences were phonotactically permissible in English but others were not. Massaro and Cohen found that the boundary between [l] and [r] varied as a function of the preceding consonant. Listeners tended to perceive [l], for example, when it was preceeded by an [s], since [#sl] is a legal sequence in English but [#sr] is not. On the other hand, [r] was favored over [l] when it followed [t] since English permits [#tr] but not [#tl].

Syntactic decisions also interact with acoustic/phonetic processes. Cooper and his colleagues (Cooper, 1980; Cooper, Paccia, & Lapointe, 1978; Cooper & Paccia-Cooper, 1980) have reported a number of instances in which rather subtle aspects of the speech signal appear to be affected by syntactic properties of the utterance. These include adjustments in the fundamental frequency, duration, and the blocking of phonological rules across certain syntactic boundaries. While these studies are concerned primarily with aspects of production, we might surmise from previous cases where perception mirrors production that listeners take advantage of such cues in perceiving speech.

Not only the accuracy, but also the speed of making low-level decisions about speech, is influenced by higher-level factors. Experimental support for this view is provided by data reported by Marslen-Wilson and Welsh (1978). In their study subjects were asked to shadow various types of sentences. Some of the utterances consisted of syntactically and semantically well-formed sentences. Other utterances were syntactically correct but semantically anomolous. A third class of utterances was both syntactically and semantically ungrammatical. Marslen-Wilson and Welsh found that shadowing latencies varied with the type of utterance. Subjects shadowed the syntactically and semantically well-formed prose most quickly. Syntactically correct but meaningless utterances were shadowed less well. Random sequences of words were shadowed most poorly of all. These results indicate that even when acoustic/phonetic analysis is possible in the absence of higher-level information, this analysis--at least as required for purposes of shadowing--seems to be aided by syntactic and semantic support.

A final example of how high-level knowledge interacts with low-level decisions comes from a study by Elman, Diehl, & Buchwald (1977). This study illustrates how phonetic categorization depends on language context ("What

language am I listening to?"). Elman et al. constructed stimulus tapes which contained a number of naturally produced one-syllable items which followed a precursor sentence. Among the items were the nonsense syllables [ba] or [pa], chosen so that several syllables had stop VOT values ranging from 0 ms. to 40 ms. (in addition to others with more extreme values).

Two tapes were prepared and presented to subjects who were bilingual in Spanish and English. On one of the tapes, the precursor sentence was "Write the word..."; the the other tape contained the Spanish translation of the same sentence. Both tapes contained the same [ba] and [pa] nonsense stimuli. Subjects listened to both tapes; for the Spanish tape in which all experimental materials and instructions were in Spanish; the English tape was heard in an English context.

The result was that subjects' perceptions of the same [ba]/[pa] stimuli varied as a function of context. In the Spanish condition, the phoneme boundary was located in a region appropriate to Spanish (i.e., near 0 ms.) while in the English condition the boundary was correct for English (near 30 ms.).

One of the useful lessons of this experiment comes from a comparison of the results with previous attempts to induce perceptual shifts in bilinguals. Earlier studies had failed to obtain such language-dependent shifts in phoneme boundary (even though bilinguals have been found to exhibit such shifts in production). Elman et al. suggested that the previous failures were due to inadequate procedures for establishing language context. These included a mismatch between context (natural speech) and experimental stimuli (synthetic speech). Contextual variables may be potent forces in perception, but the conditions under which the interactions occur may also be very precisely and narrowly defined.

Reliance on lexical constraints. Even in the absence of syntactic or semantic structure, lexical constraints exert a powerful influence on perception; words are more perceptible than nonwords (Rubin, Turvey, & VanGelder, 1976). Indeed, this word advantage is so strong that listeners may even perceive missing phonemes as present, provided the result yields a real word (Warren, 1970; Samuel, 1979). Samuel (1980) has shown that if a missing phoneme could be restored in several ways (e.g., *le_ion* could be restored either as *legion* or *lesion*), then restoration does not occur.

Speech perception occurs rapidly and in one pass. In our view, an extremely important fact about human speech perception is that it occurs in one pass and in real time. Marslen-Wilson (1975) has shown that speakers are able to shadow (repeat) prose at very short latencies (e.g., 250 ms., roughly equal to a one syllable delay). In many cases, listeners are able to recognize and begin producing a word before it has been completed. This is especially true once a portion of a word has been heard which is sufficient to uniquely determine the identity of the word. This ability of humans to process in real time stands in stark contrast to machine-based recognition systems.

Context effects get stronger toward the ends of words. Word endings appear to be more susceptible to top-down effects than word endings. Put differently, listeners appear to rely on the acoustic input less and less as more of a word is heard.

Marslen-Wilson and Welsh (1978) found that when subjects were asked to shadow prose in which errors occurred at various locations in words, the subjects tended to restore (i.e., correct) the error more often when the error occurred in the third syllable of a word (53%) than in the first syllable (45%). Cole, Jakimik, & Cooper (1978) have reported similar findings. On the other hand, if the task is changed to *error detection*, as in a study by Cole and Jakimik (1978), and we measure reaction time, we find that subjects detect errors faster in final syllables than in initial syllables.

Both sets of results are compatible with the assumption that word perception involves a narrowing of possible candidates, As the beginning of a word is heard, there may be many possibilities as to what could follow. Lack of a lexical bias would lead subjects to repeat what they hear exactly. They would also be slower in detecting errors, since they would not yet know what word was intended. As more of the word is heard, the candidates for word recognition are narrowed. In many cases, a single possibility will emerge before the end of the word has been presented. This knowledge interacts with the perceptual process so that less bottom-up information is required to confirm that the expected word was heard. In some cases, even errors may be missed. At the same time, when errors *are* detected, detection latency will be relatively fast. This is because the listener now knows what the intended word was.

PREVIOUS MODELS OF SPEECH PERCEPTION

One can distinguish two general classes of models of speech perception which have been proposed. On the one hand we find models which claim to have some psycholinguistic validity, but which are rarely specified in detail. And on the other hand are machine-based speech understanding systems; these are necessarily more explicit but do not usually claim to be psychological valid.

Psycholinguistic models. Most of the psycholinguistic models lack the kind of detail which would make it possible to test them empirically. It would be difficult, for example, to develop a computer simulation in order to see how the models would work given real speech input.

Some of the models do attempt to provide answers to the problems mentioned in the previous section. Massaro and his colleagues (Massaro & Oden, 1980a, 1980b; Oden & Massaro, 1978; Massaro & Cohen, 1977) have recognized the significance of interactions between features in speech perception. They propose that, while acoustic cues are perceived independently from one another, these cues are integrated and matched against a *propositional prototype* for each speech sound. The matching procedure involves the use of *fuzzy logic* (Zadeh, 1972). In this way their model expresses the generalization that features frequently exhibit "trading relations" with one another. The model is one of the few to be formulated in quantitative terms, and provides a good fit to the data Massaro and his co-workers have collected. However, while we value the descriptive contribution of this approach, it fails to provide an adequate statement of the mechanisms required for perception to occur.

Cole and Jakimik (1978, 1980) have also addressed many of the same concerns which have been identified here. Among other problems, they note the difficulty of segmentation, the fact that perception is sensitive to the position within a word, and that context plays an important role in speech perception. Unfortunately, their observations--while insightful and well-substantiated--have not yet led to what might be considered a real model of how the speech perceiver solves these problems.

The approach with which we find ourselves in greatest sympathy is that taken by Marslen-Wilson (Marslen-Wilson, 1975, 1980; Marslen-Wilson & Tyler, 1975; Marslen-Wilson & Welsh, 1978). Marslen-Wilson has described a model which is similar in spirit to Morton's (1979) *logogen* model and which emphasizes the parallel and interactive nature of speech perception.

In Marslen-Wilson's model, words are represented by active entities which look much like logogens. Each word element is a type of evidence-gathering entity; it searches the input for indications that it is present. These elements differ from logogens in that they are able to respond actively to mismatches in the signal. Thus, while a large class of word elements might become active at the beginning of an input, as that input continues many of the words will be