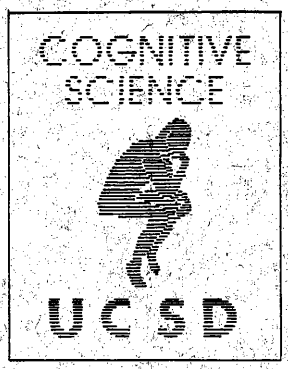
ICS REPORT NO. 8302

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.

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

disconfirmed and will remove themselves from the pool of word candidates. Eventually only a single word will remain. At this point the word is perceived. Marslen-Wilson's basic approach is attractive because it accounts for many aspects of speech perception which suggest that processing is carried out in parallel. While the model is vague or fails to address a number of important issues, it is attractive enough so that we have used it as the basis for our initial attempt to build an interactive model of speech perception. We will have more to say about this model presently.

A number of other speech perception models have been proposed, including those of Pisoni & Sawusch (1975), Cooper (1979), Liberman, Cooper, Harris, & MacNeilage (1962), and Halle & Stevens (1964), and many of these proposals provide partial solutions to the problem. For instance, while there are serious difficulties with a strong formulation of the Motor Theory of Speech Perception (Liberman et al., 1962), this theory has focused attention on an important fact. Many of the phenomena which are observed in an acoustic analysis of speech appear to be puzzling or arbitrary until one understands their articulatory foundation. There is good reason to believe that speech perception involves--if not necessarily (MacNeilage, Rootes, & Chase, 1967) at least preferably--implicit knowledge of the mapping between articulation and sound. It may well be, as some have suggested (Studdert-Kennedy, 1982) that speech perception is best understood as *event perception*, that event being speech production.

Despite insights such as these, we feel that previous models of speech perception have serious deficiencies.

First, these models are almost never formulated with sufficient detail that one can make testable predictions from them. Second, many of them simply fail to address certain critical problems. For example, few models provide any account for how the units of speech (be they phonemes, morphemes, or words) are identified given input in which unit boundaries are almost never present. Nor do most models explain how listeners are able to unravel the encoding caused by coarticulation.

While we find the greatest agreement with Marslen-Wilson's approach, there are a number of significant questions his model leaves unanswered. (1) How do the word elements know when they match the input? The failure of many machine-based speech recognition systems indicates this is far from trivial problem. (2) Do word elements have internal structure? Do they encode phonemes and morphemes? (3) How is serial order (of words, phonemes, morphemes, etc.) represented? (4) How do we recognize nonwords? Must we posit a separate mechanism, or is there some way in which the same mechanism can be used to perceive both words and nonwords? (5) How is multi-word input perceived? What happens when the input may be parsed in several ways, either as one long word or several smaller words (e.g., *sell ya light* vs. *cellulite*)? These are all important questions which are not addressed.

Machine-based models. It might seem unfair to evaluate machine-based speech recognition systems as models of speech perception, since most of them do not purport to be such. But as Norman (1980) has remarked in this context, "nothing succeeds like success." The perceived success of several of the speech understanding systems to grow out of the ARPA Speech Understanding Research project (see Klatt, 1977, for review), has had a profound influence on the field of human speech perception. As a result, several recent models have been proposed (e.g., Klatt, 1980; Newell, 1980) which *do* claim to model human speech perception, and whose use of pre-compiled knowledge and table look-up is explicitly justified by the success of the machine-based models. For these reasons, we think the machine-based systems must be considered seriously as models of human speech perception.

The two best known attempts at machine recognition of speech are **HEARSAY** and **HARPY.**

HEARSAY (Erman & Lesser, 1980; Carnegie-Mellon, 1977) was the more explicitly psychologically-oriented of the two systems. **HEARSAY** proposed several computationally distinct knowledge sources, each of which could operate on the same structured data base representing hypotheses about the contents of a temporal window of speech. Each knowledge source was supposed to work in parallel with the others, taking information from a central "blackboard" as it became available, suggesting new hypotheses, and revising the strengths of others suggested by other processing levels.

Although conceptually attractive, **HEARSAY** was not a computationally successful model (in the sense of satisfying the ARPA SUR project goals, Klatt, 1977), and there are probably a number of reasons for this. One central reason appeared to be the sheer amount of knowledge that had to be brought to bear in comprehension of utterances--even of utterances taken from a very highly constrained domain such as the specification of chess moves. Knowledge about what acoustic properties signaled which phonemes, which phonemes might occur together and how those co-occurrances condition the acoustic properties, knowledge of which sequences of speech sounds made legal words in the restricted language of the system, knowledge about syntactic and semantic constraints, and knowledge about what it made sense to say in a particular context had to be accessible. The machinery available to **HEARSAY** (and by machinery we mean the entire computational approach, not simply the hardware available) was simply not sufficient to bring all of these considerations to bear in the comprehension process in anything close to real time.

Three other problems may have been the fact that the analysis of the acoustic input rarely resulted in unambiguous identification of phonemes; the difficulties in choosing between which hypotheses would most profitably be pursued first (the "focus of attention" problem); and the fact that the program was committed to the notion that the speech input had to be segmented into separate phonemes for identification. This was a very errorful process. We will argue that this step may be unnecessary in a sufficiently parallel mechanism.

The difficulties faced by the HEARSAY project with the massive parallel computation that was required for successful speech processing were avoided by the HARPY system (Lowerre & Reddy, 1980; Carnegie-Mellon, 1977). HARPY's main advantage over HEARSAY was that the various constraints used by HEARSAY in the process of interpreting an utterance were pre-compiled into HARPY's computational structure, which was an integrated network. This meant that the extreme slowness of HEARSAY's processing could be overcome; but at the expense, it turned out, of an extremely long compilation time (over 12 hours of time on a DEC-10 computer). This trick of compiling in the knowledge, together with HARPY's incorporation of a more sophisticated acoustic analysis, and an efficient graph-searching technique for pruning the network ("beam search"), made it possible for this system to achieve the engineering goals established for it.

However, HARPY leaves us at a dead end. Its knowledge is frozen into its structure and there is no natural way for knowledge to be added or modified. It is extremely unlikely that the simplified transition network formalism underlying HARPY can actually provide an adequate formal representation of the structure of language or the flexibility of its potential use in real contexts.

* * * * *

Both the psycholinguistic and the machine models share certain fundamental assumptions about how the processing of speech is best carried out. These assumptions derive, we feel, from the belief that the van Neumann digital computer is the appropriate metaphor for information processing in the brain. This metaphor suggests that processing is carried out as a series of operations, one operation at a time; that these operations occur at high speeds; and that knowledge is stored in random locations (as in Random Access Memory) and must be retrieved through some search procedure. These properties give rise to a characteristic processing strategy consisting of iterated hypothesize-and-test loops. (It is curious that even in the case of HEARSAY, which came closest to escaping the van Neumann architecture, the designers were unwilling to abandon this fundamental strategy.)

Yet we note again how poorly this metaphor has served in developing a model for human speech perception. Let us now consider an alternative.

THE INTERACTIVE ACTIVATION MODEL OF SPEECH PERCEPTION

The Philosophy Underlying the Present Model

In contrast to **HARPY** and **HEARSAY**, we do not believe that it is reasonable to work toward a computational system which can actually process speech in real time or anything close to it. The necessary parallel computational hardware simply does not exist for this task. Rather, we believe that it will be more profitable to work on the development of parallel computational mechanisms which seem in principle to be capable of the actual task of speech perception, given sufficient elaboration in the right kind of hardware, and to explore them by running necessarily slow simulations of massively parallel systems on the available computational tools. Once we understand these computational mechanisms, they can be embodied in dedicated hardware specially designed and implemented through very large scale integration (VLSI).

Again in contrast to **HARPY** and **HEARSAY**, we wish to develop a model which is consistent with what we know about the psychology and physiology of speech perception. Of course this is sensible from a point of view of theoretical psychology. We believe it is also sensible from the point of view of designing an adequate computational mechanism. The only existing computational mechanism that can perceive speech is the human nervous system. Whatever we know about the human nervous system, both at the physiological and psychological levels, provides us with useful clues to the structure and the types of operations of one computational mechanism which is successful at speech perception.

We have already reviewed the psychological constraints, in considering reasons why the problem of speech perception is difficult and in exploring possible clues about how it occurs. In addition, there are a few things to be said about the physiological constraints.

What is known about the physiology is very little indeed, but we do know the following. The lowest level of analysis of the auditory signal is apparently a coding of the frequency spectrum present in the input. There is also evidence of some single-unit detectors in lower-order mammals for transitions in frequency either upward or downward, and some single units respond to frequency transitions away from a particular target frequency (Whitfield & Evans, 1965). Whether such single units actually correspond to functional detectors for these properties is of course highly debatable, but the sparse evidence is at least consistent with the notion that there are detectors for properties of the acoustic signal beginning at the lowest level with detectors for the particular frequencies present in the signal. Detectors may well be distributed over large populations of actual neurons, of course.

More fundamentally, we know that the brain is a highly interconnected system. The number of neurons in the cortex (conservatively, 10 billion) is not nearly as impressive as the number of synapses--perhaps as many as 10^{14}. The

connectivity of cortical cells is such that a change of state in one area is likely to influence neurons over a very wide region.

We know also that neuronal conductivity is relatively slow, compared with digital computers. Instruction cycle times of digital computers are measured on the order of nanoseconds; neuronal transmission times are measured on the order of milliseconds. Where does the power of the human brain come from, then? We suggest it derives from at least these two factors: the interconnectedness of the system, and the ability to access memories by content. Content addressable memory means that information can be accessed directly instead of accessed through a sequential scan of randomly ordered items.

This leads us toward a model which is explicitly designed to deal with all of the constraints outlined above. We have adopted the following "design principles:"

- The model should be capable of producing behavior which is as similar as possible to human speech perception. We consider experimental data to be very important in providing constraints and clues as to the model's design. The model should not only perform as well as humans, but as poorly in those areas where humans fail.

- The model should be constructed using structures and processes which are plausible given what we know about the human nervous system. We do not claim that the model is an image of those neuronal systems which are actually used in humans to perceive speech, since we know next to nothing about these mechanisms. But we have found that mechanisms which are inspired by the structure of the nervous system offer considerable promise for providing the kind of parallel information processing which seems to be necessary.

- The model should not be constrained by the requirement that computer simulations run in real time. Parallel processes can be simulated on a serial digital machine, but not at anything approaching real-time rates. The goal of real time operation at this point would be counter-productive and would lead to undesirable compromises.

The COHORT Model

Our initial attempt to construct a model which met these requirements was called the COHORT model, and it was an attempt to implement the model of that name proposed by Marslen-Wilson and Welsh (1978). Of course, in implementing the model many details had to be worked out which were not specified in the original, so the originators of the basic concept cannot be held responsible for all of the model's shortcomings. COHORT was designed to perceive word input, with the input specified in terms of time- and strength-varying distinctive features. It

is based on a lexicon of the 3846 most common words (occurring 10 or more times per million) from the Kucera & Francis corpus (Kucera & Francis, 1967).

Each of the features, phonemes, and words is represented by a *node*. Nodes have roughly the same computational power as is traditionally ascribed to a neuron. Each node has...

...an associated *level of activation* which varies over time. These levels may range from some minimum value usually near -.2 or -.3 to a maximum, usually set at +1.0;

...a *threshold* (equal to 0); when a node's activation level exceeds this threshold it enters what is called the *active* state and begins to signal its activation value to other units;

...its own (sub-threshold) *resting level* of activation to which it returns in the absence of any external inputs.

Each node may be linked to other nodes in a non-random manner. These connections may be either *excitatory* or *inhibitory*. When a node becomes active, it excites those nodes to which it has excitatory connections, and inhibits nodes to which it has inhibitory connections by an amount proportional to how strongly its activation exceeds threshold. These connections have associated weightings, such that some inputs may have relatively greater impact on a node than others.

A node's current activation level reflects several factors: (1) the node's initial resting level; (2) the spatial and temporal summation of previous inputs (excitatory and inhibitory); and (3) the node's rate of decay.

A fragment of the system just described is illustrated in Figure 1. At the lowest level we see the nodes for the acoustic/phonetic features. **COHORT** makes use of a set of 22 nodes for 11 bipolar features which are modifications of the Jakobsonian distinctive features (Jakobson, Fant, & Halle, 1952). These nodes are activated directly by the input to the model (described below). The features were chosen for the initial working model for several reasons. They have proven useful in the description of certain linguistic phenomena (such as sound change) which suggests they have some psychological reality; the Jakobsonian features are defined in (sometimes vague) acoustic terms; and recent work by Blumstein and Stevens (1980; Stevens & Blumstein, 1981) appears to confirm that some of the features might serve as models for more precise acoustic templates.

At the next higher level are the nodes for phonemes. **COHORT** has nodes for 37 different phonemes, including an abstract unit which marks the end of words. All phonemes except the end of work marker receive excitatory inputs from those features which signal their presence. Thus, the node for /*p*/ is activated by input from the nodes *GRAVE, COMPACT, CONSONANTAL, ORAL, VOICELESS*, etc.

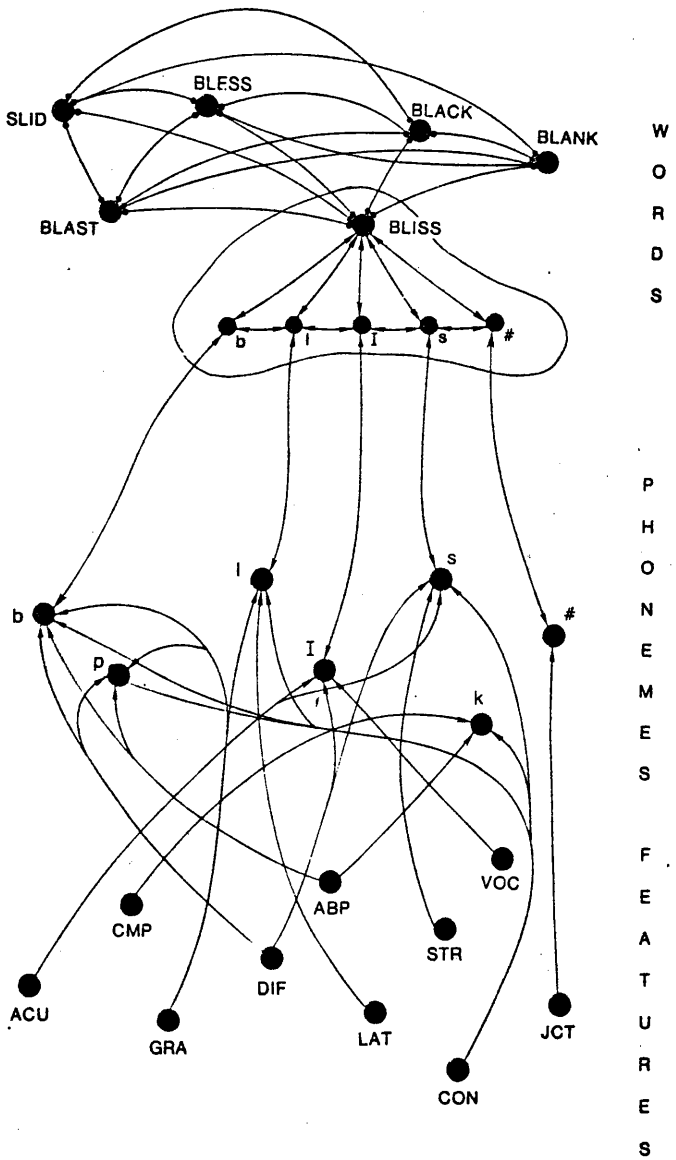


Figure 1. Fragment of the COHORT system. Nodes exist for features, phonemes, and words. The word nodes have a complex schema associated with them, shown here only for the word *bliss*. Connections between nodes are indicated by arcs; excitatory connections terminate in arrows and inhibitory connections terminate in filled circles.

Before describing the word nodes, a comment is in order regarding the features and phonemes which were used in COHORT. These choices represent initial simplifications of very complicated theoretical issues, which we have chosen not to broach at the outset. Our goal has been to treat the model as a starting place for examining a number of computational issues which face the development of adequate models of speech perception, and it is our belief that many of these issues are independent of the exact nature of the assumptions we make about the features. The Jakobsonian feature set was a convenient starting point from this point of view, but it should be clear that the features in later versions of the model will need substantial revision. The same caveat is true regarding the phonemes. It is even conceivable that some other type of unit will ultimately prove better. Again, to some degree, the precise nature of the unit (phoneme, demisyllable, context-sensitive allophone, transeme, etc.) is dissociable from the structure in which it is embedded.

It might be argued that other choices of units would simplify the problem of speech perception considerably and make it unnecessary to invoke the complex computational mechanisms we will be discussing below. Indeed, some of the units which have been proposed as alternatives to phonemes have been suggested as answers to the problem of context-sensitive variation. That is, they encode--frozen into their definition--variations which are due to context. For example, context-sensitive allophones (Wickelgren, 1969) attempt to capture differences the the realizations of particular phonemes in different contexts by imagining that there is a different unit for each different context. We think this merely postpones a problem which is pervasive throughout speech perception. In point of fact, none of these alternatives is able to truly solve the variability which extends over broad contexts, or which is due to speaker differences, or to changes in rate of articulation. For this reason we decided to begin with units (phonemes) which are frankly context-insensitive, and to see if their variability in the speech stream could be dealt with through the processing structures.

Let us turn now to the word nodes. Words present a special problem for COHORT. This is because words contain internal structure. In the current version of the system, this structure is limited to phonemes, but it is quite likely that word structure also contains information about morphemes and possibly syllable boundaries. To account for the fact that words are made up of ordered sequences of phonemes, it seems reasonable to assume that the perceiver's knowledge of words specifies this sequence.

Word nodes are thus complex structures. A node network which depicts a word structure is shown for the word *bliss* in Figure 1. The schema consists of several nodes, one for each of the phonemes in the word, and one for the word itself. The former are called token nodes, since there is one for each occurrance of each phoneme in the word. The latter is simply called the word node. At the end of each word there is a special token node corresponding to a word boundary.

Token nodes have several types of connections. *Token-word* connections permit tokens to excite their word node as they become active (pass threshold). *Word-token* links allow the word node to excite its constituent tokens. This serves both to reinforce tokens which may have already received bottom-up input, as well as to prime tokens that have not yet been "heard." *Phoneme-token* connections provide bottom-up activation for tokens from phonemes. Finally, *token-token* connections let active tokens prime successive tokens and keep previous tokens active after their bottom-up input has disappeared. Because listeners have some expectation that new input will match word beginnings, the first token node of each word has a slightly higher resting level than the other tokens. (In some simulations, we have also set the second token node to an intermediate level, lower than the first and higher than the remaining tokens). Once the first token passes threshold, it excites the next token in the word. This priming, combined with the order in which the input actually occurs, is what permits the system to respond differently to the word *pot* than to *top*.

In addition to internal connections with their token nodes, word nodes have inhibitory connections with all other word nodes. This inhibition reflects competition between word candidates. Words which match the input will compete with other words which do not, and will drive their activation levels down.

Word recognition in COHORT

To further illustrate how COHORT works, we will describe what is involved in recognizing the word *slender*.

COHORT does not currently have the capability for extracting features from real speech, so we must provide it with a hand-constructed approximation of those features which would be present in the word *slender*. Also, since the model is simulated on a digital computer, time is represented as a series of discrete samples. During each sampling period COHORT receives a list of those features which might be present during that portion of the word. These features have time-varying strengths. To simulate one aspect of coarticulation, the features overlap and rise and fall in strength.

At the beginning of the simulation, all nodes in the system are at their resting levels. During the first few sampling periods the feature nodes receive activation from the input, but their activation levels remain below threshold. Eventually, however, some feature nodes become active and begin to excite all the phonemes which contain them. In the present example, activation of the features for /s/ results in excitation of the phonemes /z/, /f/, and /v/ as well as /s/. This is because these other phonemes closely resemble /s/ and contain many of the same features. The /s/, however, is most strongly activated.

The next thing that happens is that active phonemes excite their corresponding token nodes in all the words that contain those phonemes. Initial token nodes (such as the /s/ in *slender*) are more likely to pass threshold than

word-internal nodes (such as the /s/ in *twist*) since these nodes have higher resting levels. When the token nodes become active, they begin to activate word nodes and also their successor token nodes. Of course, while all this happens, input continues to provide bottom-up excitation.

As time goes on, the internal connections begin to play an increasing role in determining the state of the system. Once word nodes become active they provide a strong source of top-down excitation for their token nodes and also compete with one another via inhibitory connections. Early in the input there may be many words which match the input and are activated. These will compete with one another but none will be able to dominate; however, they will drive down the activations of other words. Those words which fail to continue to receive bottom-up excitation will fall away, both through their own decay and through inhibition from more successful candidates. Eventually only a single word will remain active and will push down the activation levels of unsuccessful word nodes.

One can monitor these events by examining the activation levels of the various types of nodes in the system. In Figure 2, for example, we see a graph of the activation levels of word nodes, given input appropriate to the word *slender*. At time t_0 the word nodes' activation levels rest just below threshold. During the first 15 or so time cycles the activation levels remain constant, since it takes a while for the feature, phoneme, and token nodes to become active and excite the word nodes. After this happens a large number of words become active. These are all the words which begin with the phoneme /s/. Shortly after the 25th cycle features for the phoneme /l/ are detected and words such as *send* fall away, but other words such as *slim* remain active. When the /e/ is detected *slim* and similar words are inhibited. At the end only *slender* remains active.

This simulation reveals two interesting properties of COHORT. First, we note that occasionally new words such as *lend* and *endless* join the cohort of active words. Even though they do not begin with /s/ they resemble the input enough to reach threshold. We regard this as desirable because it is clear that human listeners are able to recover from initial errors. One problem we have found in other simulations is that COHORT does not display this behavior consistently enough.

Secondly, we see that the word node for *slender* begins to dominate surprisingly early in time. In fact, it begins to dominate at just the point where it provides a unique match to the input. This agrees with Marslen-Wilson's (1980) claim that words are recognized at the point where they become uniquely identifiable.

We can also monitor the activation levels of the tokens within the word schema for *slender*, as shown in Figure 3. At time t_0 all tokens are below threshold, although /s/ is near threshold and the /l/ is also slightly higher than the remaining tokens. (Recall that the initial tokens have higher resting levels, reflecting perceiver's expectations for hearing first sounds first.) The /s/ token

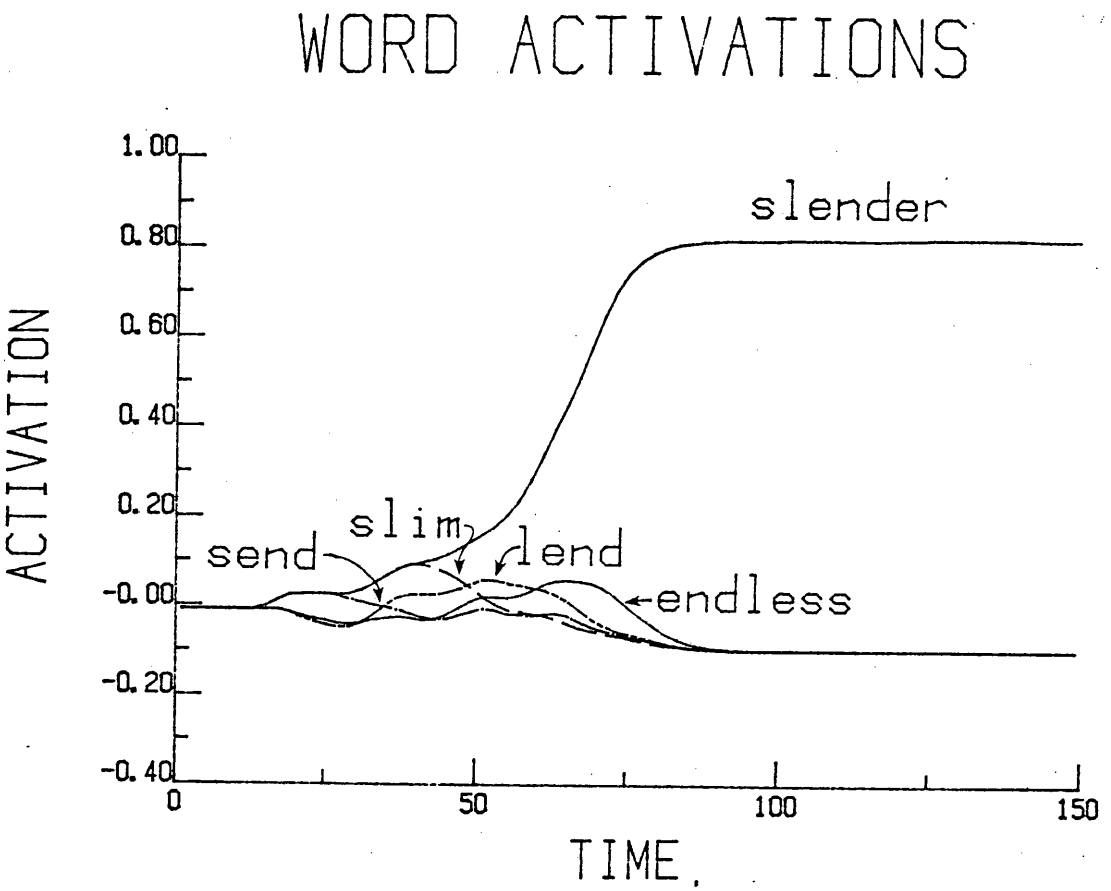


Figure 2. Activation levels of selected word nodes, given feature inputs appropriate for the word *slender*. At the start all words which begin with *s* are activated. As time goes on only those words which more closely resemble the input remain active; other words are decay and are also inhibited by the active nodes. Finally only the node for *slender* dominates.

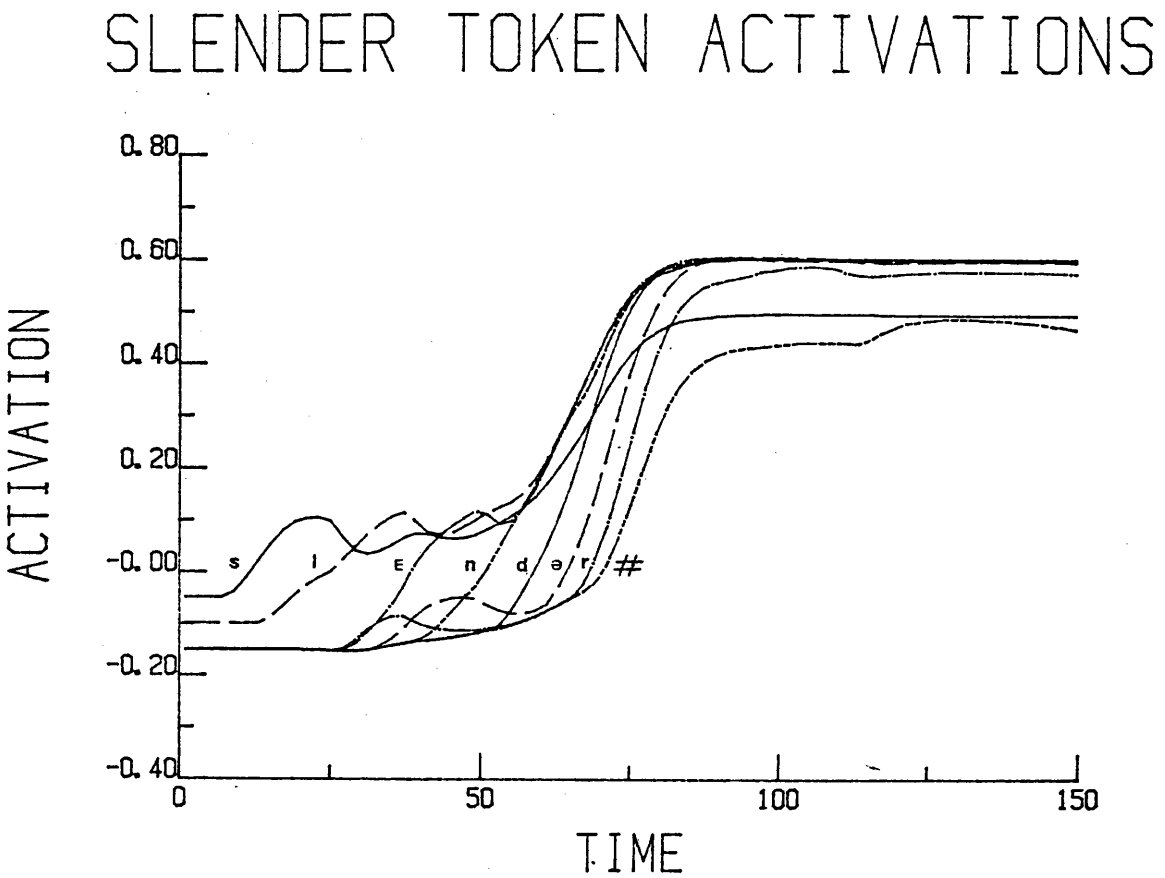


Figure 3. Activations of the token nodes associated with *slender*, given input appropriate for this word.

passes threshold fairly quickly. When it becomes active it excites both the *slender* word node and also the next token in the word, /l/. After more cycles, the /l/ token begins to receive bottom-up input from the feature nodes, and the /s/'s feature input decreases.

The same basic pattern continues throughout the rest of the word, with some differences. The level of nodes rises slowly even before they receive bottom-up input and become active. This occurs because the nodes are receiving lateral priming from earlier tokens in the word, and because once the word node becomes active it primes all its constituent token nodes. This lateral and top-down excitation is also responsible for the tendency of token nodes to increase again after decaying once bottom-up input has ceased (for example, /s/'s level starts to decay at cycle 25, then begins to increase at cycle 30). By the end of the word, all the tokens are very active, despite the absence of any bottom-up excitation.

This example demonstrates how COHORT deals with two of the problems we noted in the first section. One of these problems, it will be recalled, is the spreading of features which occurs as a result of coarticulation. At any single moment in time, the signal may contain features not only of the "current" phoneme but also neighboring phonemes. In the current version of COHORT we provide the simulation with hand-constructed input in which this feature spreading is artificially mimicked. Because COHORT is able to activate many features and phonemes at the same time, this coarticulation helps the model anticipate phonemes which may not, properly speaking, be fully present. In this way coarticulation is treated as an aid to perception, rather than as a source of noise. While the sort of artificial input we provide obviously does not provide the same level of difficulty which is present in real speech, we believe that COHORT's approach to dealing with these rudimentary aspects of coarticulation is on the right track.

A second problem faced by many speech recognition systems is that of segmentation: How do you locate units in a signal which contains few obvious unit boundaries? For COHORT this problem simply never arises. As the evidence for different phonemes waxes and wanes, the activation levels of phonemes and tokens rises and falls in continuous fashion. Tokens which are activated in the right sequence (i.e., belong to real words) activate word nodes, which are then able to provide an additional source of excitation for the tokens. At the end of the process, all the phoneme tokens of the word that has been heard are active, but there is no stage during which explicit segmentation occurs.

In addition to these two characteristics, COHORT can be made to simulate two phenomena which have been observed experimentally in human speech perception. The first of these phenomena is phonemic restoration.

The human speech processing system is capable of perceiving speech in the face of considerable noise. This ability was studied in an experiment by Warren (1970). Warren asked subjects to listen to tapes containing naturally produced words in which portions of the words had been replaced by noise. Warren found

that, although subjects were aware of the presence of noise, they were unaware that any part of the original word had been deleted (in fact, they were usually unable to say where in the word the noise occurred). Samuel (in press) has replicated and extended these using a signal detection paradigm. (In Samuel's experiments, some stimuli have phonemes *replaced* by noise and other stimuli have noise *added in*. The subjects' task is to determine whether the phoneme is present or absent.) One of Samuel's important findings is that this phenomenon, phonemic restoration, actually completes the percept so strongly that it makes subjects insensitive to the distinction between the replacement of a phoneme by noise and the mere addition of noise to an intact speech signal. Listeners actually perceive the missing phonemes as if they were present.

We were interested in seeing how COHORT would respond to stimuli in which phonemes were missing. To do this, we prepared input protocols in which we turned off feature input during those cycles which corresponded in time to a particular phoneme. In one of these simulations, we deleted all feature input for the /d/ of *slender*. (Note that this differs slightly from the standard phonemic restoration experiment, in which noise is added to the signal after a phoneme is deleted.)

In Figure 4 we observe the activations of the *slender* token nodes which result from this input. These levels may be compared with those in Figure 3. There are no obvious differences between the two conditions. The /d/ token succeeds in becoming active despite the absence of bottom-up input. This suggests that the token-token priming and the top-down excitation from word to token is a powerful force during perception.

Figure 5 compares the word node activation for *slender* with and without /d/ input. The two patterns are remarkably alike. COHORT appears to respond much as human perceivers do given similar input -- the distinction between the presence and the absence of the /d/ is lost in context.

A second phenomenon we attempted to replicate with COHORT was the lexical bias in phoneme identification first noted by Ganong (1980). As previously mentioned, Ganong discovered that if listeners are asked to identify the initial consonant in stimuli which range perceptually from a word to a nonword, the phoneme boundary is displaced toward the word end of the continuum, compared with its location on a non-word/word continuum. In short, lexical status affects perception at the level of phonetic categorization.

In order to simulate this experiment, we presented COHORT with input which corresponded to a word-initial bilabial stop, followed by features for the sequence *_ar*. The feature values for the bilabial stop were adjusted in such a way as to make it indeterminate for voicing; it sounded midway between *bar* and *par*. Although COHORT knows the word *bar*, it does not have *par* in its lexicon, so *par* is effectively a nonword for the purposes of the simulation.

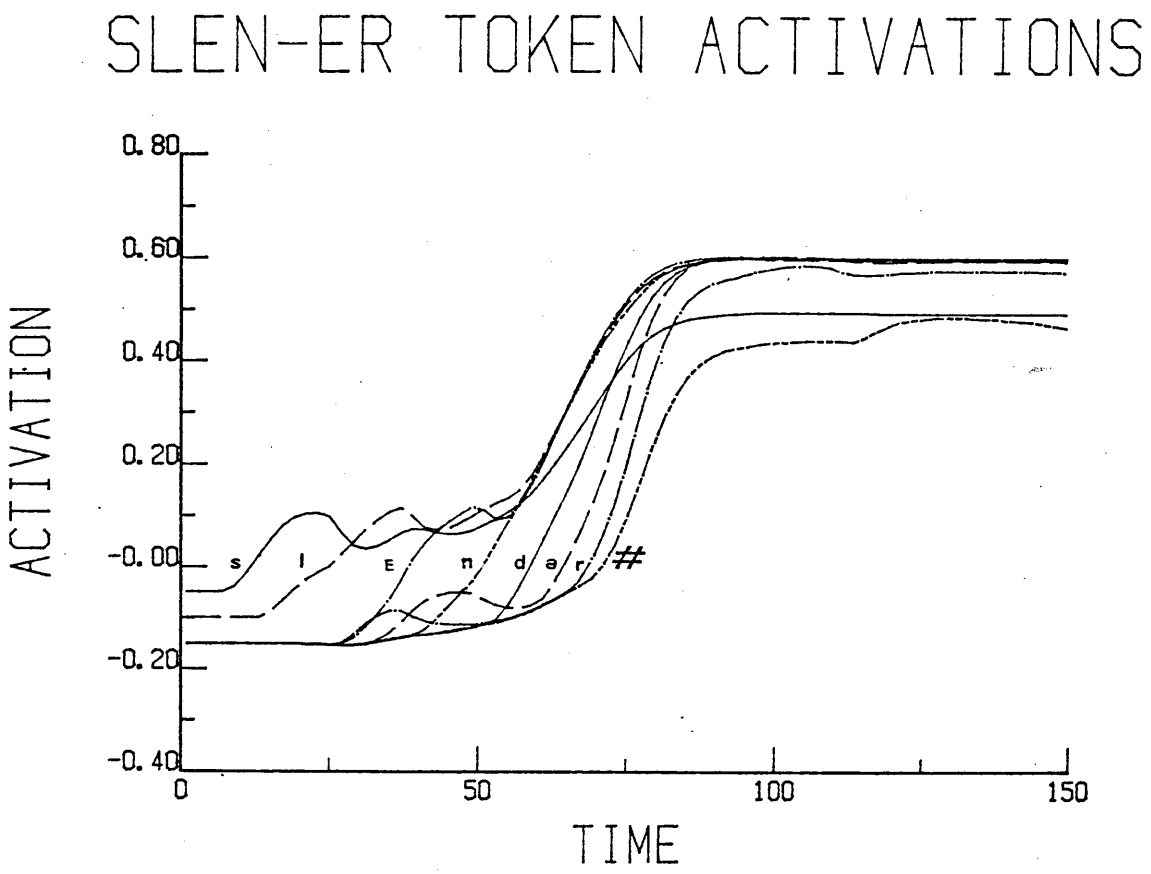


Figure 4. Activations of the token nodes associated with *slender*, given input appropriate for this word.

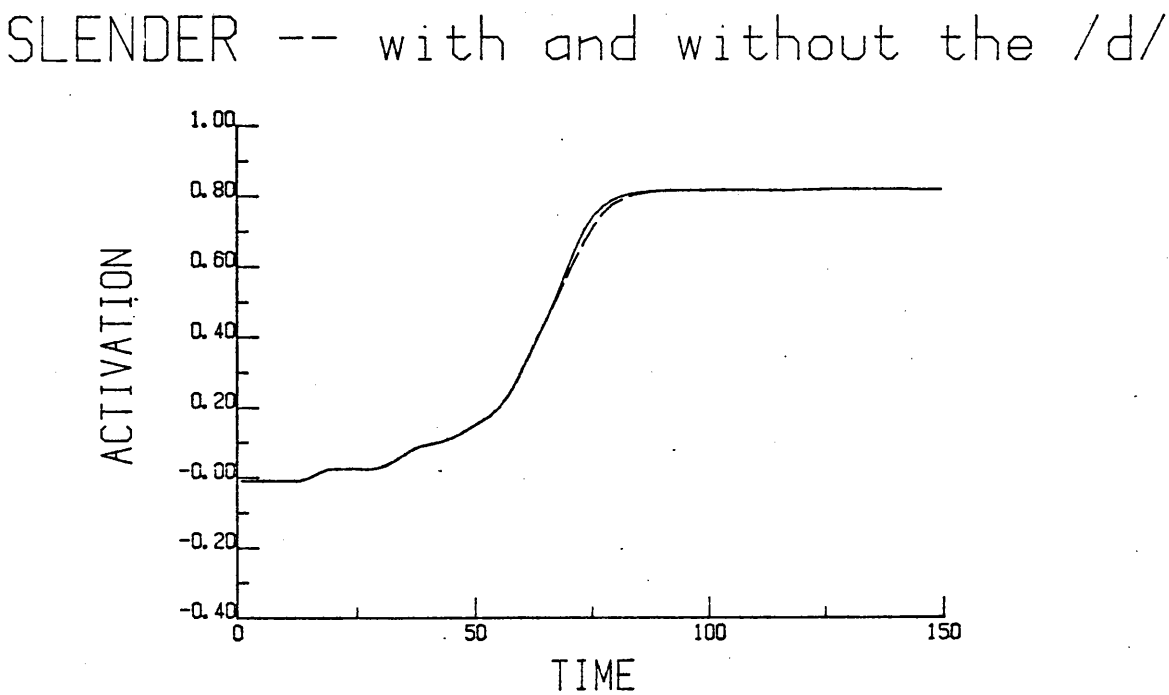


Figure 5. Activation levels of the *slender* word node for input in which the *d* is present (solid line), compared to when the *d* is absent (broken line).

The simulation differed from Ganong's experiment in that he measured the phoneme boundary shift by presenting a series of stimuli to subjects and then calculating the boundary as the location of the 50% labelling crossover. In our experiment we were able to present the model with a stimulus which should have been exactly at the phoneme boundary, assuming a neutral context (e.g., if the stimulus had been a nonsense syllable such as *ba* or *pa* rather than a potential word). The way we determined whether or not a lexical effect similar to Ganong's had occurred was to examine the activation levels of the /b/ and /p/ phoneme nodes.

Figure 6 shows the activation levels of these two nodes over the time course of processing the input stimulus. Both nodes become highly activated during the first part of the word. This is the time when bottom-up input is providing equal activation for both voiced and voiceless bilabial stops. Once the bottom-up input is gone, both levels decay. What is of interest is that the /b/ node remains with a higher level of activation. We assume that this higher level would be reflected in a boundary shift on an phoneme identification test toward the voiced end of the continuum.

When we think about why COHORT displays this behavior--behavior which is similar to those of Ganong's human subjects--we realize that the factors responsible for the greater activation of the /b/ node are essentially the same which cause phonemic restoration. Top-down excitation from the word level exerts a strong influence on perception at the phoneme level.

This realization leads to an interesting prediction. Because the lexical effect reflects the contribution of top-down information, it should be the case that when the target phoneme (i.e., the one to be identified) occurs later in the word, rather than at the beginning as is the case with the *bar/par* stimulus., the difference in activations of the two competing nodes should be magnified. This is because the word node has had longer to build up its own activation and is therefore able to provide greater support for the phoneme which is consistent with it.

Figure 7 demonstrates that COHORT does indeed perform in this manner. We presented the simulation with input appropriate to the sequence *ro_*followed by a bilabial stop that was again intermediate with regard to voicing. *rob* is a word in COHORT's lexicon, but *rop* is not, so we would expect a greater level in activation for /b/ than for /p/, based on top-down excitation.

This indeed occurs. But what we also find is that the magnitude of the difference is slightly greater than when the target phoneme occurs at the beginning of the word. The finding has not yet been tested with human perceivers, but it is consistent with other findings mentioned above (Cole & Jakimik, 1978, 1980; Marslen-Wilson & Welsh, 1978) which point to greater top-down effects at word endings than at word-beginnings.

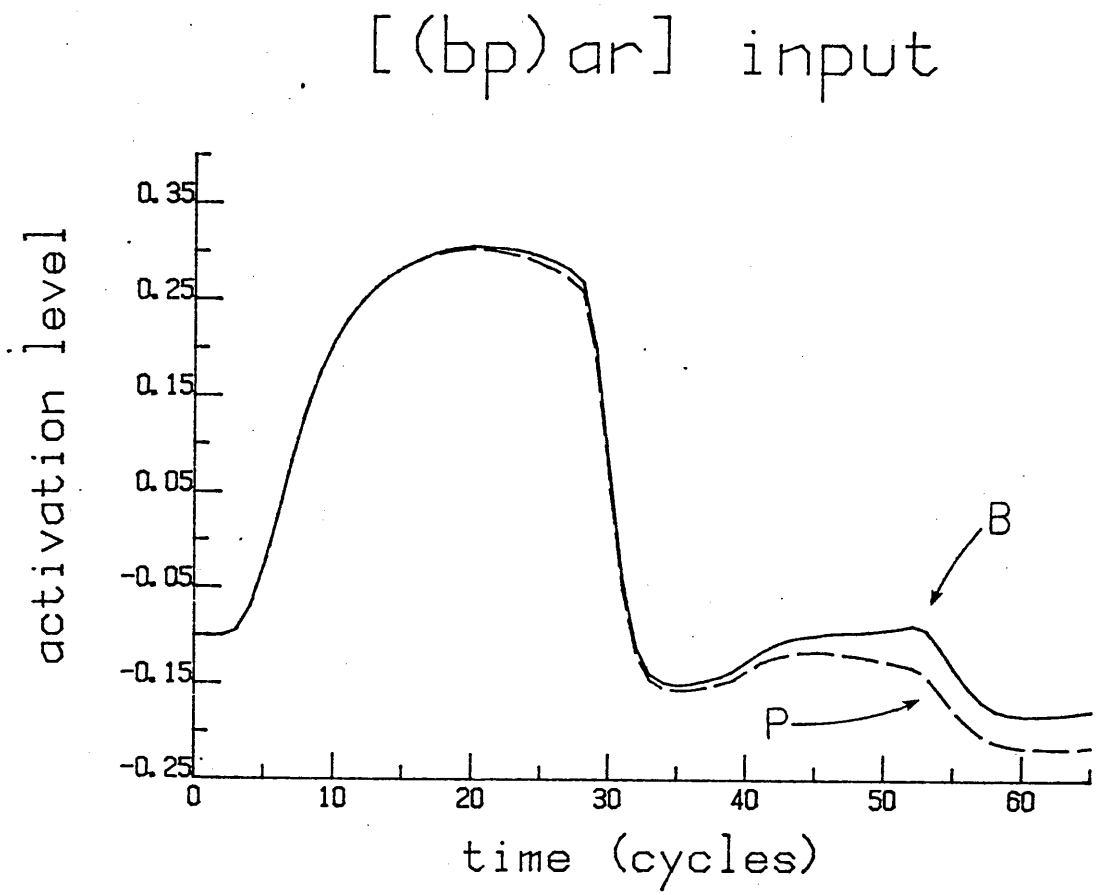


Figure 6. Activation of *b* and *p* phoneme nodes, given feature input for the sequence *bilabial stop+a+r*, in which the stop is indeterminate for voicing. Since the lexicon contains the word *bar* but not *par*, top-down excitation favors the perception of the stop as voiced.

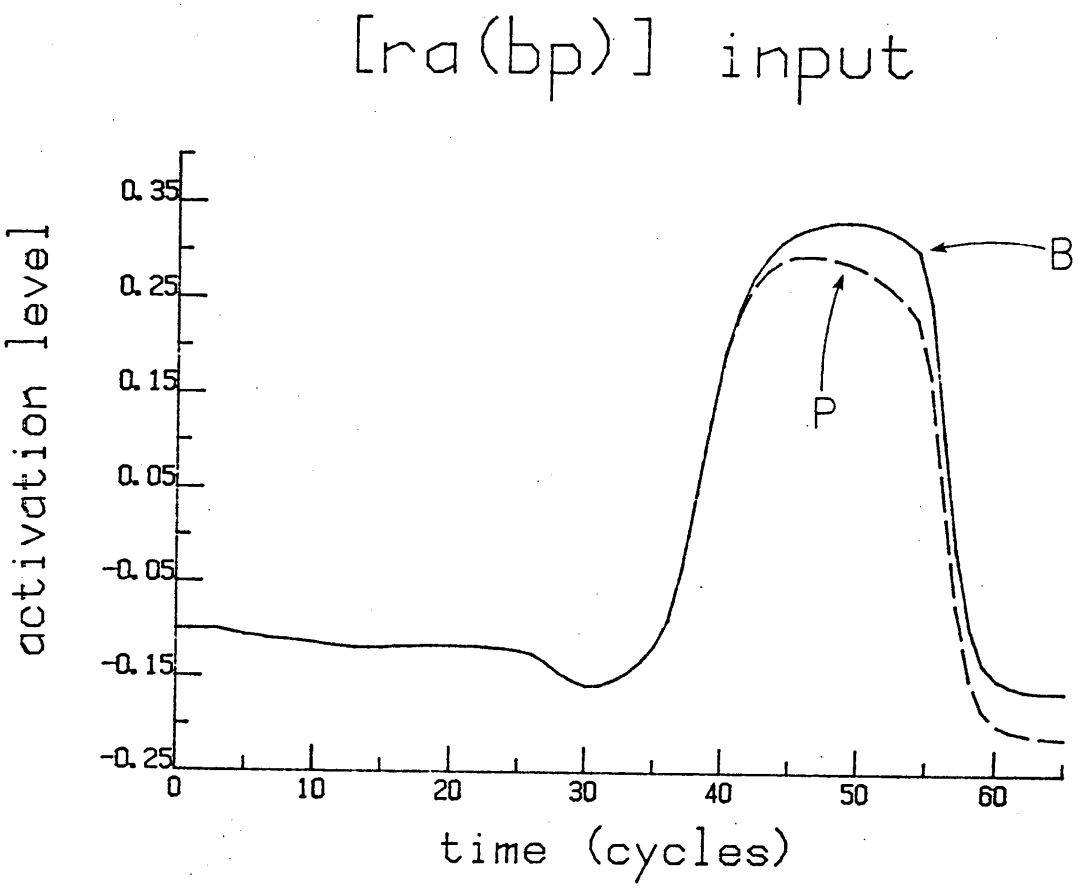


Figure 7. Activation of *b* and *p* phoneme nodes, given feature input for the sequence *r+a+bilabial stop*, in which the stop is indeterminate for voicing. The lexicon contains the word *rob*, but not *rop*, so the *b* node becomes more activated than the *p* node.

In simulating Ganong's lexical effect on the phoneme boundary, we added a provision to COHORT which was not provided for by Marslen-Wilson and Welsh (1978): Feedback from the word to the phoneme level. They, along with Morton (1979) have accounted for lexical and other contextual effects on phoneme identification in terms of a two step process, in which context affects word identification, and then the phonological structure of the word is unpacked to determine what phonemes it contains.

The alternative we prefer is to permit feedback from the words to actually influence activations at the phoneme level. In this way, partial activations of words can influence perception of nonwords.

The addition of feedback from the words to the phoneme level in cohort raises a serious problem, however. If the feedback is strong enough so that the phoneme nodes within a word are kept active as the perceptual process unfolds, then all words sharing the phonemes which have been presented continue to receive bottom-up support and the model begins to loose its ability to distinguish words having the same phonemes in them in different orders. This and other problems, to be reviewed below, have lead us to a different version of an interactive activation model of speech perception, called TRACE.

The TRACE Model

Given COHORT's successes, one might be tempted to suggest that it may be feedback to the phoneme level, and not the rest of the assumptions of COHORT which are in error. However, there are other problems as well with this version of the model. First, words containing multiple occurrances of the same phoneme present serious problems for the model. The first occurrance of the phoneme primes all the tokens of this phoneme in words containing this phoneme anywhere. Then the second occurrance pushes the activations of all of these tokens into the active range. The result is that words containing the repeated phoneme anywhere in the word become active. At the same time, all words containing multiple occurrances of the twice-active phoneme get so strongly activated that the model's ability to distinguish between them based on subsequent (or prior) input is diminished. A second difficulty is that the model is too sensitive to the durations of successive phonemes. When durations are too short they do not allow for sufficient priming. When they are too long too much priming occurs and the words being to "run away" independently of bottom-up activation.

In essence, both of these problems come down to the fact that COHORT uses a trick to handle the sequential structure of words: it uses lateral priming of one token by another to prepare to perceive the second phoneme in a word after the first and so on. The problems described above arise from the fact that this is a highly unreliable way of solving the problem of the sequential structure of speech. To handle this problem there needs to be some better way of directing

the input to the appropriate place in the word.

Sweeping the input across the tokens. One way to handle some of these problems is to assume that the input is sequentially directed to the successive tokens of each word. Instead of successive priming of one token by the next, we could imagine that when a token becomes active, it causes subsequent input to be gated to its successor rather than itself. All input, of course, could be directed initially toward the first token of each word. If this token becomes active, it could cause the input to be redirected toward the next token. This suggestion has the interesting property that it automatically avoids double activation of the same token on the second presentation of the corresponding phoneme. It may still be sensitive to rate variations, though this could be less of a problem than in the preceeding model. Within word filling in could still occur via the top-down feedback from the word node, and of course this would take a while to build up so would be more likely to occur for later phonemes than for earlier ones.

However, this scheme shares a serious problem with the previous one. In the absence of prior context, both versions depend critically on clear word beginnings to get the right word schemas started. We suspect that it is inferior to human perceivers in this respect. That is, we suspect that humans are able to recognize words correctly from their endings (in so far as these are unique) even when the beginnings are sufficiently noisy so that they would produce only very weak word-level activations at first and thus would not get the ball rolling through the word tokens.

Generalized sweeping. A potential solution to this problem would be to sweep the input through all tokens, not just those in which the input has already produced activations. However, it is not clear on what basis to proceed with the sweep. If it were possible to segment the input into phonemes then one could step along as each successive phoneme came in; but we have argued that there is no segmentation into phonemes. Another possibility is to step along to the next token as tokens become active at the current position in any words. Though this does not require explicit segmentation of the input, it has its drawbacks as well. For one thing it means that the model is somewhat rigidly committed to its position within a word. It would be difficult to handle cases where a nonsense beginning was followed by a real word (as in, say, unpticohort), since the model would be directing the ending toward the ends of longer words rather than toward beginnings.

The memory trace. A problem with all of the schemes considered thus far is that they have no memory, except within each word token. Patterns of activation at the phoneme level come and go very quickly -- if they do not, confusion sets in. The fact that the memory is all contained within the activations of the word tokens makes it hard to account for context effects in the perception of pseudowords (Samuel, 1979). Even when these stimuli are not recognized as words, missing phonemes which are predictable on the basis of regularities in patterns of phoneme co-occurrance are nevertheless filled in. Such phenomena suggest that

there is a way of retaining a sequence of phonemes -- and even filling in missing pieces of it -- when that sequence does not form a word. One possibility is to imagine that the activations at the phoneme level are read out into some sort of post-identification buffer as they become active at the phoneme level. While this may account for some of the pseudoword phenomena, retrospective filling in of missing segments would be difficult to arrange. What appears to be needed is a dynamic memory in which incomplete portions of past inputs can be filled in as the information which specifies them becomes available. The TRACE model attempts to incorporate such dynamic memory into an interactive activation system. We are only now in the process of implementing this model via a computer simulation, so we can only offer the following sketch of how it will work.

We propose that speech perception takes place within a system which possesses a dynamic representational space which serves much the same function as the Blackboard in **HEARSAY.** We might visualize this buffer as a large set of banks of detectors for phonetic features and phonemes, and imagine that the input sweeps out a pattern of activation through this buffer. That is, the input at some initial time t_0 would be directed to the first bank of detectors, the input at the next time slice would be directed to the next bank, and so on. These banks are dynamic; that is, they contain nodes which interact with each other, so that processing will continue in them after bottom-up input has ceased. In addition to the interactions within a time slice, nodes would interact across slices. Detectors for mutually incompatible units would be mutually inhibitory, and detectors for the units representing an item spanning several slices would support each other across slices. We assume in this model that information written into a bank would tend to decay, but that the rate of decay would be determined by how strongly the incoming speech pattern set up mutually supportive patterns of activation within the trace.

Above the phoneme model, we presume that there would be detectors for words. These, of course, would span several slices of the buffer. It seems unreasonable to suppose that there is an existing node network present containing nodes for each word at each possible starting position in the buffer. It seems then, that the model requires the capability of creating such nodes when it needs them, as the input comes in. Such nodes, once created, would be interact with the phoneme buffers in such a way as to insure that only the correct sequence of phonemes will strongly activate them. Thus, the created node for the word *cat* starting in some slice will be activated when there is a /c/ in the starting slice and a few subsequent slices, an /a/ in the next few slices, and a /t/ in the next few, but will not be excited (except for the /a/) when these phonemes occur in the reverse order.

A simplified picture of the **TRACE** model is shown in Figure 8. Time is represented along the horizontal axis, with successive columns for individual memory traces. Within each trace there are nodes for features and phonemes, but only phoneme nodes are shown here. The activation level of each of these nodes (and of the word nodes above) is shown as a horizontal bar; thicker bars indicate greater levels of activation.

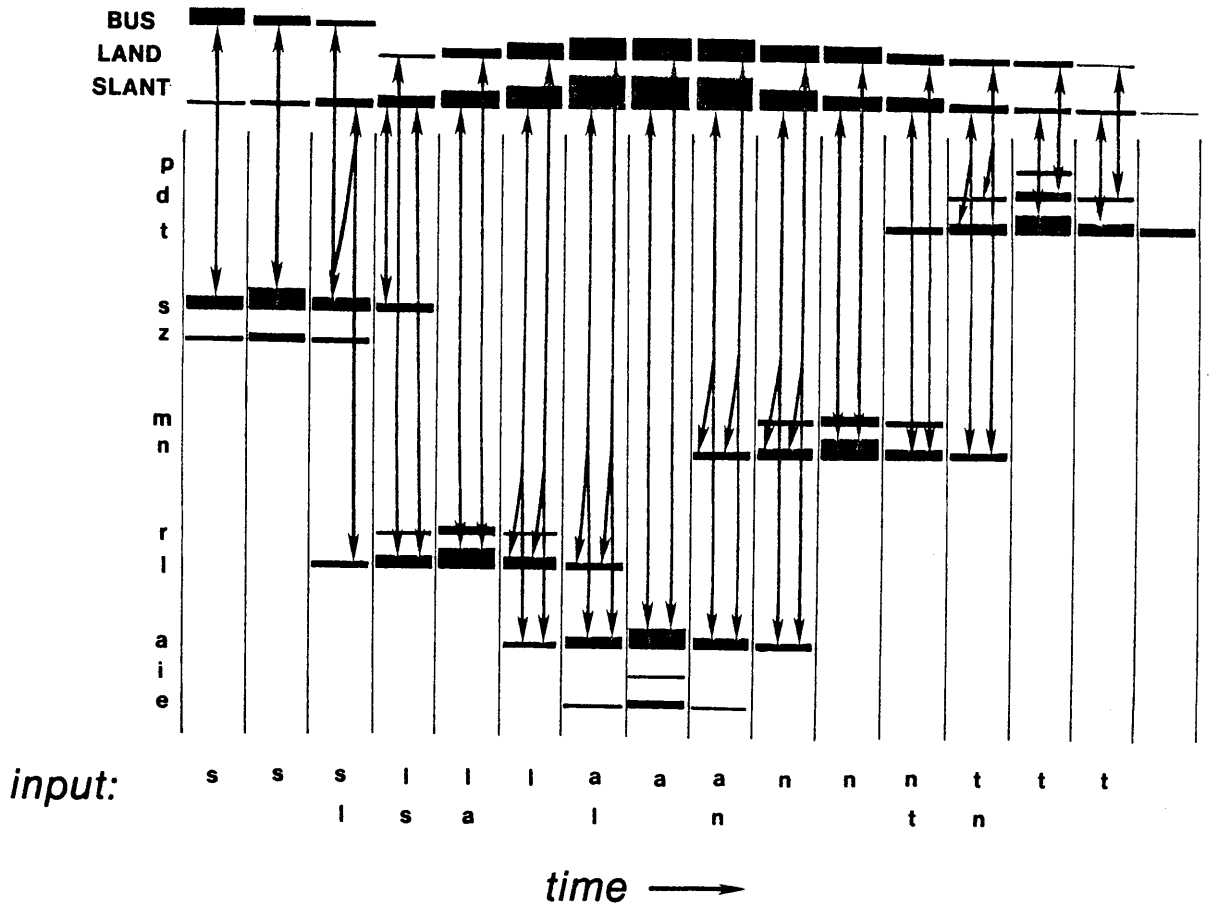


Figure 8. Partial view of the TRACE system. Time is represented along the horizontal axis, with columns for succeeding "traces." Each trace contains nodes for phoneme and feature nodes (only the phoneme nodes are shown). Input is shown along the bottom in phonemic form; in reality, input to the phoneme nodes would consist of excitation from the feature nodes within each trace. At the top are shown the word nodes and the activations they receive in each time slice. Because the input can be parsed in various ways, several word nodes are active simultaneously and overlap.

Along the bottom is shown sample input. The input is presented here in phonemic form for ease of representation; it would actually consist of the excitations from the (missing) feature nodes, which in turn would be excited by the speech input.

Because the input as shown could be parsed in different ways, the word nodes for *slant*, *land*, and *bus* all receive some activation. *slant* is most heavily activated since it most closely matches the input, but the sequence *bus land* is also entertained. Presumably context and higher-level information are used to provide the necessary input to disambiguate the situation.

In this model, we can account for filling-in effects in terms of top-down activations of phonemes at particular locations in the trace. One important advantage of the TRACE model is that a number of word tokens partially consistent with a stretch of the trace and each weakly activating a particular phoneme could conspire together to fill in a particular phoneme. Thus if the model heard *fluggy*, words which begin with *flu...* such as *fluster* and *flunk* would activate phoneme nodes for /f/, /l/, and /ə/ in the first part of the trace, and words which end with *...uggy* such as *buggy* and *muggy* would activate nodes for /ʌ/, /g/, and /i/ in the latter part of the trace. In this way the model could be made to account easily for filling in effects in pseudoword as well as word perception.

This mechanism for using the lexicon to perceive non-words is intriguing, because it suggests that some of the knowledge which linguists have assumed is represented by rules might located in the lexicon Instead. Consider, for example, phonotactic knowledge. Every language has certain sequences of sounds which are permissible and others which are not. English has no word *blik*, but it might, whereas most speakers of English would reject *bnik* as being unacceptable. One might choose to conclude, therefore, that speakers have rules of the form

* #bn

(where the asterisk denotes ungrammaticality, and # indicates word beginning), or more generally

* #[stop] [nasal]

But in fact, TRACE points to an alternative account for this behavior. If perception of both words and nonwords is mediated by the lexicon, then to the extent that a sequence of phonemes in a nonword occurs in the real words, TRACE will be able to sustain the pattern in the phoneme traces. If a sequence does not exist, the pattern will still be present in the trace, but only by virtue of bottom-up input, and weakly. TRACE predicts that phonotactic knowledge may not be hard-and-fast in the fashion that rule-governed behavior should be. Because there are some sequences which are uncommon, but which do occur in English (e.g., initial *sf* clusters) listeners should be able to judge certain nonwords as more acceptable than others; and this is in fact what happens (Greenberg & Jenkins, 1964).

Another advantage to **TRACE** is that early portions of words would still be present in the trace and so would remain available for consideration and modification. Ambiguous early portions of a word could be filled in retrospectively once subsequent portions correctly specified the word. This would explain listeners' tendencies to hear an [h] in the phrase *_eel of the shoe* (Warren & Sherman, 1974).

The **TRACE** model permits more ready extension of the interactive activation approach to the perception of multi-word input. One can imagine the difficulties which would be presented in **COHORT** given input which could be parsed either as a single word, or several smaller words. Consider, for example, what would happen if the system heard a string which could be interpreted either as *sell ya' light* or *cellulite*. Assume that later input will disambiguate the parsing, and that for the time being we wish to keep both possibilities active. Because words compete strongly with one another in **COHORT**, the nodes for *sell*, *your*, *light*, and *cellulite*, will all be in active competition with one another. The system will have no way of knowing that the competition is really only between the first three of these words--as a group--and the last. In **TRACE**, words still compete, but the competition can be directed toward the portion of the input they are attempting to account for.

CONCLUSIONS

That speech perception is a complex behavior is a claim which is hardly novel to us. What we hope to have accomplished here is to have shed some light about exactly what it is about speech perception which makes it such a difficult task to model, and to have shown why interactive activation models are such an appropriate framework for speech perception. Our basic premise is that attempts to model this area of human behavior have been seriously hampered by the lack of an adequate computational framework.

During the course of an utterance a large number of factors interact and shape the speech stream. While there may be some acoustic invariance in the signal, such invariance seems to be atypical and limited. It seems clear that attempting to untangle these interactions within human information processing frameworks which resemble von Neumann machines is a formidable task. Those computer-based systems which have had any success, such as **HARPY**, have achieved real-time performance at the expense of flexibility and extensibility, and within a tightly constrained syntactic and lexical domain. We do not wish to downplay the importance of such systems. There are certainly many applications where they are very useful, and by illustrating how far the so-called "engineering" approach can be pushed they provide an important theoretical function as well.

However, we do not believe that the approach inherent in such systems will ever lead to a speech understanding system which performs nearly as well as humans, at anywhere near the rates we are accustomed to perceiving speech. There is a fundamental flaw in the assumption that speech perception is carried out in a processor which looks at all like a digital computer. Instead, a more adequate model of speech perception assumes that perception is carried out over a large number of neuron-like processing elements in which there are extensive interactions. Such a model makes sense in terms of theoretical psychology; we would argue that it will ultimately prove to be superior in practical terms as well.

In this chapter we have described the computer simulation of one version (**COHORT**) of an interactive activation model of speech perception. This model reproduces several phenomena which we know occur in human speech perception. It provides an account for how knowledge can be accessed in parallel, and how a large number of knowledge elements in a system can interact. It suggests one method by which some aspects of the encoding due to coarticulation might be decoded. And it demonstrates the paradoxical feat of extracting segments from the speech stream without ever doing segmentation.

COHORT has a number defects. We have presented an outline of another model, **TRACE**, which attempts to correct some of these defects. **TRACE** shows that it is possible to integrate a dynamic working memory into an interactive activation model, and that this not only provides a means for perceiving nonwords

but also shows that certain type of knowledge can be stored in the lexicon which leads to what looks like rule-governed behavior.

What we have said so far about TRACE is only its beginning. For one thing, the process by which acoustic/phonetic features are extracted from the signal remains a challenging task for the future. And we have yet to specify how the knowledge above the word level should come into play. One can imagine schema which correspond to phrases, and which have complex structures somewhat like words, but there are doubtless many possibilities to explore.

It is clear that a working model of speech perception which functions anywhere nearly as well as humans do is a long way off. We do not claim that any of the versions we present here are the right ones, but we are encouraged by the limited success of COHORT and the potential we see in TRACE. The basic approach is promising.

ACKOWLEDGEMENTS

The research reported here was funded by a contract with the Office of Naval Research (M00014-82-C-0374), 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.

REFERENCES

Alfonso, P. J. Context effects on the perception of place of articulation. Paper presented to the meeting of the Acoustical Society of America, Ottowa, Canada, May 1981.

Blumstein, S. E., & Stevens, K. N. Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, 1980, 67, 648-662.

Bolt, Beranek, & Newman. *Speech understanding systems: Final Technical Progress Report.* BBN Report No. 3438, 1976.

Carnegie-Mellon University, *Speech understanding systems: Summary of the five-year research effort at Carnegie-Mellon University*, Department of Computer Science, CMU, 1977.

Chuang, C.-K., & Wang, W. S.-Y. Psychophysical pitch biases related to vowel quality, intensity difference, and sequential order. *Journal of the Acoustical Society of America*, 1978, *64*, 1004-1014.

Cole, R.A., & Jakimik, J. Understanding speech: How words are heard. In G. Underwood (Ed.), *Strategies of information processing*. London: Academic Press, 1978, pp. 149-211.

Cole, R.A., & Jakimik, J. A model of speech perception. In Cole, R.A. (Ed.), *Perception and production of fluent speech.* Hillsdale, N.J.: Erlbaum, 1980, pp. 133-163.

Cole, R., Jakimik, J., & Cooper, W. E., Perceptibility of phonetic features in fluent speech. *Journal of the Acoustical Society of America.* 1978, *64*, 44-56.

Cooper, W. E. *Speech perception and production.* Norwood, N.J.: Ablex, 1979.

Cooper, W. E. Syntactic-to-phonetic coding. In B. Butterworth (Ed.), *Language production.* New York: Academic Press, 1980, pp. 211-298.

Cooper, W. E., Paccia, J. M., & LaPointe, S. G. Hierarchical coding in speech timing. *Cognitive Psychology*, 1978, *10*, 154-177.

Cooper, W. E., & Paccia-Cooper, J. *Syntax and speech.* Cambridge, Mass.: Harvard University Press, 1980.

Derr, M. A., & Massaro, D. W. The contribution of vowel duration, F_0 contour, and frication duration as cues to the /juz/-/jus/ distinction. *Perception & Psychophysics*, 1980, 51-59.

Dorman, M. F., Raphael, L. J., & Liberman, A. M. Further observations on the role of silence in the perception of stop consonants. *Journal of the Acoustical Society of America*, 1976, *59*, S40.

Elman, J.L., Diehl, R.L., & Buchwald, S.E. Perceptual switching in bilinguals. *Journal of the Acoustical Society of America*, 1977, *62*, 971-974.

Erman, L.D., & Lesser, U.R. The Hearsay-II speech understanding system: A tutorial. In W.A. Lea, *Trends in speech recognition*. Englewood Cliffs, N.J.: Prentice-Hall, 1980, pp. 361-381.

Fujimura, O., & Lovins, J. B. Syllables as concatenative phonetic units. In A. Bell and J. B. Hooper (Eds.), *Syllables and segments*. Amsterdam: North-Holland, 1978, pp. 107-120.

Ganong, W.F. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 1980, *6*, 110-125.

Greenberg, J.H., & Jenkins, J.J. Studies in the psychological correlates of the sound system of American English. *Word*, 1964, *20*, 157-177.

Halle, M., & Stevens, K. N. Speech recognition: A model and a program for research. In J. A. Fodor & J. J. Katz (Eds.), *The structure of language: Readings in the psychology of language*. Englewood Cliffs, N. J.: Prentice-Hall, 1964.

Harris, K.S. Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech*, 1958, *1*, 1-17.

Hasegawa, A. *Some perceptual consequences of fricative coarticulation*. Unpublished doctoral dissertation. Purdue, 1976.

Hinton, G. E., & Anderson, J. A. (Eds.) Parallel models of associative memory. Hillsdale, N. J.: Erlbaum, 1981.

Isenberg, D., Walker, E.C.T., Ryder, J.M., & Schweikert, J. A top-down effect on the identification of function words. Paper presented at the Acoustical Society of America, Los Angeles, November 1980.

Jakobson, R. *Child language, aphasia, and phonological universals*. The Hague: Mouton, 1968.

Jakobson, R., Fant, G., & Halle, M. *Preliminaries to Speech Analysis*. Cambridge : MIT Press, 1952.

Klatt, D.H. Review of the ARPA Speech Understanding Project. *Journal of the Acoustical Society of America*, 1977, *62*, 1345-1366.

Klatt, D.H. Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. Cole (Ed.), *Perception and production of fluent speech.* Hillsdale, N.J.: Erlbaum, 1980, pp. 243-288.

Kucera, H., & Francis, W. *Computational analysis of present-day American English.* Providence, R. I.: Brown University Press, 1967.

Liberman, A. M., Cooper, F. S., Harris, K. S., & MacNeilage, P. F. A motor theory of speech perception. *Proc. Speech Communication Seminar.* Stockholm: Royal Institute of Technology, 1962.

Liberman, A.M., Cooper, F.S., Shankweiler, D., & Studdert-Kennedy, M. Perception of the speech code. *Psychology Review*, 1967, *84*, 452-471.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 1952, *65*, 497-516.

Liberman, A. M., Delattre, P. C., Gerstman, L. J., & Cooper, F. S. Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology*, 1956, *52*, 127-137.

Lindblom, B.E.F. Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 1963, *35*, 1773-1781.

Lindblom, B.E.F., & Studdert-Kennedy, M. On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, 1967, *42*, 830-843.

Lisker, L. Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, 1957, *33*, 42-49.

Lisker, L. *Rapid* vs. *rabid:* A catalogue of acoustic features that may cue the distinction, *Status Report of Speech Research SR-54.* New Haven, Conn.: Haskins Laboratories, 1978.

Lowerre, B., and Reddy, R. The Harpy Speech Understanding System. In Wayne A. Lea (Ed.), *Trends in Speech Recognition.* Englewood Cliffs, N.J.: Prentice-Hall, 1980, pp. 340-360.

MacNeilage, P. F., Rootes, T. P., & Chase, R. A. Speech production and perception in a patient with severe impairment of somesthetic perception and motor control. *Journal of Speech and Hearing Research*, 1967, *10*, 449-467.

Mann, V. A. Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, 1980, *28*, 407-412.

Mann, V. A., & Repp, B. H. Influence of vocalic context on perception of the [š]-

[s] distinction. *Perception & Psychophysics*, 1980, *28*, 213-228.

Marslen-Wilson, W. D. Sentence perception as an interactive parallel process. *Science*, 1975, *189*, 226-228.

Marslen-Wilson, W. D. Speech understanding as a psychological process. In Simon, J. C. (Ed.), *Spoken Language Generation and Understanding*, New York: Reidel, 1980, pp. 39-67.

Marslen-Wilson, W. D., & Tyler, L. K. Processing structure of sentence perception. *Nature*, 1975, *257*, 784-786.

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

Martin, J.G., & Bunnell, H.T. Perception of anticipatory coarticulation on effects. *Journal of the Acoustical Society of America*, 1981, *69*, 559-567.

Massaro, D. W., & Cohen, M. M. The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. *Journal of the Acoustical Society of America*, 1976, *60*, 704-717.

Massaro, D. W., & Cohen, M. M. The contribution of voice-onset time and fundamental frequency as cues to the /zi/-/si/ distinction. *Perception & Psychophysics*, 1977, *22*, 373-382.

Massaro, D. W., & Cohen, M. M. Phonological constraints in speech perception. Paper presented to the meeting of the Acoustical Society of America, Atlanta, Georgia, April 1980.

Massaro, D. W., & Oden, G. C. Speech perception: A framework for research and theory. In N. Lass (Ed.), *Speech and language: Advances in basic research and practice, Vol. 3*. New York: Academic Press, 1980, pp. 129-165.

Massaro, D. W., & Oden, G. C. Evaluation and integration of acoustic features in speech perception. *Journal of the Acoustical Society of America*, 1980, *67*, 996-1013.

McClelland, J. L., & Rumelhart, D. E. An interactive activation model of context effects in letter perception, Part I: An account of basic findings. *Psychological Review*, 1981, *375-407*.

Miller, J. L. Effects of speaking rate on segmental distinctions. In P. D. Eimas and J. L. Miller (Eds.), *Perspectives on the study of speech*. Hillsdale, N. J.: L. Erlbaum, 1981, pp. 39-74.

Miller, J. L., & Eimas, P. D. Studies on the perception of place and manner of

articulation: A comparison of the labial-alveolar and nasal-stop distinctions. *Journal of the Acoustical Society of America*, 1977, *61*, 835-845.

Miller, J.L., & Liberman, A.M. Some effects of late-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, 1979, *25*, 457-465.

Minifie, F., Kuhl, P., & Stecher, B. Categorical perception of [b] and [w] during changes in rate of utterance. *Journal of the Acoustical Society of America*, 1976, *62*, S79.

Morton, J. Word recognition. In J. Morton and J. C. Marshall (Eds.), *Psycholinguistics 2: Structures and Processes*. Cambridge, Mass.: M.I.T. Press, 1979, pp. 107-156.

Newell, A. Harpy, production systems, and human cognition. In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, N.J.: L. Erlbaum, 1980, pp. 289-380.

Norman, D.A. Copycat Science or Does the mind really work by table look-up? In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, N.J.: L. Erlbaum, 1980, pp. 381-395.

Oden, G. C., & Massaro, D. W. Integration of featural information in speech perception. *Psychological Review*, 1978, *85*, 172-191.

Ohman, S. E. G. Coarticulation in VCV utterances. *Journal of the Acoustical Society of America*, 1966, *34*, 151-168.

Peterson, G. E., & Barney, H. L. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 1952, *24*, 175-184.

Pisoni, D. B. In defense of segmental representations in speech processing. Paper presented to the Acoustical Society of America, Ottowa, Canada, May 1981.

Pisoni, D. B., & Sawusch, J. R. Some stages of processing in speech perception. In A. Cohen and S. G. Nooteboom (Eds.), *Structure and process in speech perception*. New York: Springer-Verlag, 1975, pp. 16-35.

Repp, B. H. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Haskins Status Reports on Speech Research*, 1981, *SR-67/68*, 1-40.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of acoustic cues for stop, fricative and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance, 1978, 4*, 621-637.

Repp, B. H., & Mann, V. A. Perceptual assessment of fricative-stop coarticulation. *Journal of the Acoustical Society of America*, 1980, *69*, 1154-1163.

Repp, B. H., & Mann, V. A. Fricative-stop coarticulation: Acoustic and perceptual evidence. *Haskins Laboratories Status Report on Speech Research*, 1981, *SR-67/68*, 255-266.

Rubin, P.E., Turvey, M.T., & VanGelder, P. Initial phonemes are deleted faster in spoken words than n spoken nonwords. *Perception & Psychophysics*, 1976, *19*, 394-398.

Rumelhart, D. E., & McClelland, J. L. An interactive activation model of context effects in letter perception, Part II: The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, 1981, in press.

Samuel, A. G. *Speech is specialized, not special.* Unpublished Ph. D. Thesis, University of California, San Diego, 1979.

Samuel, A. G. The effect of lexical uniqueness on phonemic restoration. Paper presented to the meeting of the Acoustical Society of America, Los Angeles, November 1980.

Samuel, A. G. Phonemic restoration: Insights from a new methodoloy. *Journal of Experimental Psychology: General*, 1981, in press.

Stevens, K., & Blumstein, S. The search for invariant acoustic correlates of phonetic features. In P.D. Eimas and J.L. Miller (Eds.), *Perspectives on the Study of Speech.* Hillsdale, N.J.: Erlbaum, 1981, pp. 1-38.

Strange, W., Verbrugge, R. R., & Shankweiler, D. P. Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, 1976, *60*, 213-224.

Strevens, P. Spectra of fricative noise. *Language and Speech*, 1960, *3*, 32-49.

Studdert-Kennedy, M. Speech perception. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics.* New York: Academic Press, 1976, pp. 243-293.

Studdert-Kennedy, M. The emergence of phonetic structure. *Haskins Laboratories Status Report on Speech Research*, 1982, *SR-69*, 217-222.

Summerfield, A. Q. Towards a detailed model for the perception of voicing contrasts. *Speech Perception, No. 3.*, Department of Psychology, Queen's University of Belfast, 1974.

Verbrugge, R. R., & Shankweiler, D. Prosodic information for vowel identity.

Journal of the Acoustical Society of America, 1977, *61*, S39.

Verbrugge, R. R., Shankweiler, D., & Fowler, C. Context-coordinated specification of vowel identity. In J. J. Wolf and D. H. Klatt (Eds.), *Speech communication papers.* New York: Acoustical Society of America, 1976.

Warren, R. M. Perceptual restoration of missing speech sounds. *Science*, 1970, *167*, 393-395.

Warren, R.M., & Sherman, G. Phonemic restorations based on subsequent context. *Perception & Psychophysics*, 1974, *16*, 150-156.

Whitefield, J.C., & Evans, E.F. Responses of auditory cortical neurons to stimuli of changing frequency. *Journal of Neurophysiology*, 1965, *28*, 655-672.

Wickelgren, W.A. Context-sensitive coding, associative memory and serial order in (speech) behavior. *Psychological Review*, 1969, *76*, 1-15.

Zadeh, L. A. A fuzzy-set-theoretic interpretation of linguistic hedges. *Journal of Cybernetics*, 1972, *2*, 4-34.

Zue, V. W. *Acoustic characteristics of stop consonants: A controlled study.* Technical Report No. 523, Lincoln Laboratory (M.I.T.), 1976.

Zue, V.W., & Schwartz, R.M. Acoustic processing and phonetic analysis. In W.A. Lea (Ed.), *Trends in speech recognition.* Englewood Cliffs, N.J.: Prentice-Hall, 1980, pp. 101-124.