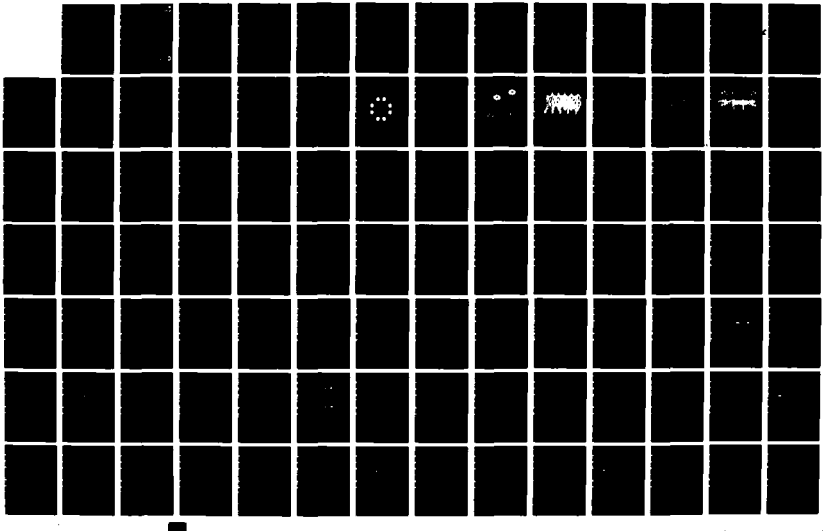
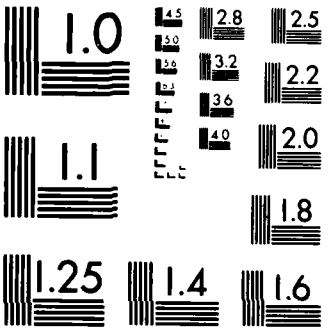
AD-A148 593 INTERACTIVE ACTIVATION MODEL OF SPEECH PERCEPTION(U) 1/2
CALIFORNIA UNIV SAN DIEGO LA JOLLA
J L MCCLELLAND ET AL. 01 NOV 84 N00014-82-C-0374
UNCLASSIFIED F/G 17/2 NL





ONR FINAL STATUS REPORT

INTERACTIVE ACTIVATION MODEL OF SPEECH PERCEPTION

Drs. James L. McClelland & Jeffrey L. Elman

Date: 1 November 1984

Period Covered by this Report: 1 March 1982 to September 30 1984

Contract Number: N00014-82-C-0374

Contract Authority Identification Number: NR 667-483

Effective Date of Contract: 1 March 1982

Expiration Date of Contract: 30 September 1984

Amount of Contract: $300,000

Scientific Officers: Dr. Henry Halff, Dr. Susan Chipman, Dr. Michael Shafto

Contractor: The Regents of University of California

Principal Investigators:

James L. McClelland
Phone: (412) 578-2789

Jeffrey L. Elman
Phone: (619) 452-2536

Sponsored by
Office of Naval Research

1. Research Program Summary

This summary is presented in two parts. The first part follows. It provides a brief overview of the goals and results of the contract research. The second part appears as a more detailed review of the research and is appended to this document.

1.1. Objectives

The objective of this research was to construct a model of human speech perception, using an interactive activation framework. Computer simulations of several versions of this model were developed. One version accepted real speech as input and attempted to A second version took simulated speech input and extracted phonemes, words, and simple phrases. The model was able to account for a variety of phenomena which have been observed in human speech perception; it also suggested behavior which had not previously been observed in human listeners but which was subsequently verified through experimental work.

1.1.1. The COHORT Model

The first version of our interactive activation was called **COHORT**, in recognition of the influence which William Marslen-Wilson's model (by that name) had on our thinking.

COHORT was comprised of a number of processing elements with a number of interconnections. Elements--called nodes--represented hypotheses about the speech input. There were three different types of nodes: distinctive feature nodes, phoneme nodes, and word nodes. Nodes were connected in ways which were consistent with the ways in which the hypotheses themselves either supported or disagreed with one another.

COHORT's input consisted of a time-varying array of numbers which corresponded to the strength of various phonetic feature. These numbers were themselves simulated by the program in such a way as to roughly corresponded with the coarticulation of features in real speech.

The "output" of the model consisted of the activation levels attained by each of the nodes. Greater activation levels corresponded to high probability of an elements existence in the input.

This model was able to simulate a variety of phenomena which have been observed in human listeners, including (1) the ability to reconstruct missing phoneme from a word ("phoneme restoration"); (2) the ability to bias phonemic perception, given an ambiguous phoneme, in a way which results in a word being perceived (as opposed to a non-word).

However, the model lacked the ability to represent a very important aspect of speech: namely, the fact that the units ordered in time. The linkage between the units perceived and the time domain was simply ignored in our system. This resulted in identical perception for mirror-image inputs (e.g., "pin" vs. "nip"). There were other problems as well, such as the overly-simplistic definition of acoustic phonetic features.

1.1.2. The TRACE Model

These deficiencies led us to an alternative approach, in which time was explicitly represented. We called this system the **TRACE** Model.

The TRACE model resembled COHORT in many important respects. The two-dimensional structure was quite similar. That is, there existed nodes for features, phonemes, and words. Nodes were connected with inhibitory or excitatory links, as before.

TRACE differed from COHORT in having a third dimension: time. TRACE possessed a memory, which enabled it to bind time with the speech input. We took the radical step of making our memory an integral part of the processing system. Indeed, in TRACE there is no essential difference between the structures which serve memory and those which serve immediate (current) perception.

The way this was accomplished was to replicate the processing network across time. Each time slice had a network of nodes which was responsible for processing that moment in time. As time progressed, the speech input was directed towards successive time networks (we called these TRACES).

This system allowed us to preserve the serial order. More significantly, it allowed the same kinds of interactions between different knowledge sources (if we consider a node to be a

knowledge source) that occurred within a time slices to occur across time.

TRACE significantly elaborated the processing structures contained in COHORT. The distinctive feature system was more sophisticated and in fact eventually allowed us to process real speech input. TRACE embodied a number of contextual interactions which affected how sounds affected the pronunciation of adjacent sounds. This made it possible to take exploit the variability in the speech signal when that variability was the result of coarticulatory effects. TRACE also contained a much richer vocabulary. The representation of time as a perceptual dimension allowed us to explore interactions at the level of word-processing, so that we could begin to tackle the problem of how words are "pulled out" of a continuous speech stream.

The current version of TRACE has the following properties:

(1) The model is able to recognize abstract segments in an input which contains no explicit segment boundaries.

(2) Memory is treated as not different than the mechanisms which carry out "current" perception, rather than consisting of passive storage buffers.

(3) The model demonstrates that the enormous variability in the speech signal which arises from contextual variability can be treated as a rich source of information, rather than as an undesirable source of noise.

(4) The lexicon provides a top-down source of input to perception. This accounts for the lexical biases in speech perception, and also for the differential status of word beginnings and word endings.

(5) Finally, the model demonstrates that at least one type of behavior which is traditionally described by linguists as rule-governed can be induced in the model without recourse to explicit rules.

A detailed description of TRACE is contained in the attached document, which will be issued as a separate Technical Report.

2. Key Personnel

Thomas Ward. Programmer: April 1982-April 1983. Ward assisted in the development of the signal processing software and design of the front-end feature extraction routines.

Paul Smith. Programmer Lab Assistant: April 1983-February 1984. Smith helped develop graphics routines for displaying results of simulations.

Mark Johnson. Research Assistant: June 1984-August 1984. Johnson helped develop graphics software and ran experiments.

David Pare. Programmer: July 1984-September 1984. Pare developed routines for creating, setting up, and accessing data structures in pilot work on a programmable version of the TRACE model.

3. Summary of Substantive Information Derived from Special Events

None.

4. Problems Encountered and/or Anticipated

None.

5. Action Required by the Government

None.

6. Fiscal Status

1. Amount provided on contract: $300,000
2. Expenditures and commitments to date: $300,000
3. Funds required to complete work: 0

7. Publications

The following publications resulted from the contract.

Elman, J.L., & McClelland, J.L. Speech perception as a cognitive process: The interactive activation model of speech perception. In Norman Lass (Ed.), *Language and Speech*. New York: Academic Press. 1984.

Elman, J.L., & McClelland, J.L. Exploiting lawful variability in the speech wave. In J.S. Perkell and D.H. Klatt (Eds.). *Invariance and Variability of Speech Processes*

Hillsdale, N.J.: Lawrence Erlbaum Associates Inc. In press.

Elman, J.L., & McClelland, J.L. An architecture for parallel processing in speech recognition: The TRACE model. In M.R. Schroeder (Ed.), *Speech Recognition*. Gottingen: Biblioteca Phonetica. In press.

McClelland, J.L., & Elman, J.L. The TRACE model of speech perception. Submitted for publication, *Cognitive Psychology*.

ABSTRACT

We describe a model called the TRACE model of speech perception. The model is based on the principles of interactive activation; that is, information processing takes place via the excitatory and inhibitory interactions of a large number of simple processing units, each working continuously to update its own activation on the basis of the activations of other nodes to which it is connected. The model is instantiated in two simulation programs. The first, TRACE I, deals with short segments of real speech, and illustrates how the principles of interactive activation allow us to exploit the lawful variability inherent in the speech wave, and to identify phonemes without relying on prior segmentation of the speech stream into a sequence of separate segments. The second simulation program, TRACE II, deals with "simulated" speech consisting of sequences of overlapping phoneme-specifications. At the phoneme level, TRACE II is used to show how the principles of interactive activation provide a natural account of the trading relations among speech cues and of categorical perception. In addition, TRACE II shows how lexical information can influence the identification of phonemes, and accounts for the fact that lexical effects are found under certain conditions but not others. The model also shows how knowledge of phonological constraints can be embodied in particular lexical items, but can still be used to influence processing of novel, non-word utterances. At the lexical level, TRACE II shows how interactive activation mechanisms capture the major positive features of Marslen-Wilson's COHORT model of speech perception, while at the same time overcoming some of its computational limitations. Finally, TRACE II illustrates how lexical information can be used to segment a stream of speech into a sequence of words and to find word beginnings and endings. Though the TRACE model has some limitations, several of which are discussed, it represents a step toward a psychologically and computationally adequate model of the process of speech perception.

Consider the perception of the phoneme /g/ in the sentence "She received a valuable gift." There are a large number of cues in this sentence to the identity of this phoneme. Acoustic information in the location of the /g/ provides one important source of cues, but there are many others. The immediate phonetic environment of the /g/ imposes lawful constraints on the acoustic realization of the /g/, and the context must be processed if these constraints are to be properly taken into account. The word in which the /g/ occurs provides another source of cues, for if we know the rest of the phonemes in this word, there are only a few possibilities for what can fill the word-initial slot. The semantic and syntactic context further constrain the possible words which might occur, and thus limit still further the possible identities of the word initial phoneme. There is ample evidence that all of these different sources of constraint are used in speech perception (see Elman and McClelland, 1984, for a review).

In this paper, we describe a model of the process of speech perception. The model is constructed within a framework which appears to be ideal for the exploitation of simultaneous, and often mutual, constraints. This framework is the interactive activation framework (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1981, 1982). According to the interactive-activation approach, information processing takes place through the excitatory and inhibitory interactions among a large number of processing elements called nodes. Each node is a very simple processing device. It stands for a hypothesis about the input being processed at a particular level of analysis and in a particular role or location in the input. The activation of a node is monotonically related to the strength of the hypothesis for which the node stands. Nodes which are mutually consistent are mutually excitatory, and nodes that are mutually inconsistent are mutually inhibitory. Thus, the node for /g/ at some time t in a speech stream has mutually excitatory interactions with nodes for words beginning with /g/ at t, and has mutually inhibitory connections with nodes for other phonemes at t.

In an interactive activation model, we imagine that all the nodes are continually receiving input from all the nodes which send them excitatory and inhibitory signals, and are continually updating their activation on the basis of these inputs. This continuous process is simulated as a series of discrete time steps, in which first all the influences of each node on every other node are accumulated, and then the activations of all the nodes are updated in preparation for the next cycle of interactions. The details of this process will be specified below.

Criteria and Constraints on Model Development

There are generally two kinds of models of the speech perception process. One kind of model, which grows out of Speech Engineering and Artificial Intelligence, attempts to provide a machine solution to the problem of speech perception. A second kind of model, growing out of Cognitive Psychology, attempts to account for psychological data on the perception of speech.

Each approach honors a different criterion for success. Machine models are judged in terms of actual performance in recognizing real speech. Psychological models are judged in terms of their ability to account for details of human performance in speech recognition. We call these two criteria *computational* and *psychological* adequacy.

Our approach has been to try to build a model which is sensitive to both of these criteria at once. This is, of course, no simple task, and we do not claim complete success on either score. But we do feel that this approach is a fruitful one. Each kind of criterion imposes its own constraints on the model. The attempt to find a way to honor both at once leads, we think, to models which will ultimately be more satisfactory judged by either criterion separately. After all, the human perceptual system is the best machine we have for recognizing speech, and it seems likely that details of its behavior will offer clues to the development of more powerful machine models. At the same time, the struggle to develop a model which achieves real performance criteria with real speech input imposes

constraints on psychological theorizing which are often overlooked in psychological models that strive primarily to account for specific aspects of psychological performance as revealed in experiments.

The model we will describe in this paper is not, of course, fully adequate, either computationally or psychologically. However, our work has been sensitive to these two kinds of constraints. We have, therefore, developed a version of the simulation program that can accept real speech -- albeit from very simple utterances at this time -- as well as a version which is computationally simpler but allows us to illustrate both how our approach can overcome certain computational limitations of other approaches and how it can account for a number of important phenomena in the experimental literature on human speech perception.

Before we turn to a description of the model itself, we begin with a brief consideration of some of the computational challenges posed by speech and of some of the psychological phenomena which any model of speech perception must account for. These computational challenges and psychological phenomena will be the principle foci of our development of the model.

Computational challenges.

Exploiting lawful variability. Any model that takes seriously the criterion of computational adequacy must be sensitive to the overwhelming variability of the speech signal. The acoustic signal representing a particular phoneme varies with the ambient acoustic environment; with the speaker's sex, dialect, vocal tract, and individual style; with the ongoing rate of speaking and the prosodic contour of the sentence; and with the local phonetic and morphological context.

While some psychological theories of speech perception place great stress on some of these influences (e.g., the motor theory of speech perception, Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967), there are few detailed models of the specific mechanisms whereby a hearer

copes with all of this variability. Many machine models, on the other hand, simply treat most of these sources of variability as noise. But the variability is generally lawful, and the psychological evidence to be cited below indicates that human listeners do exploit it. Thus, the first computational challenge which we have set ourselves is to find ways to *exploit the lawful variability* in the speech wave.

Representation without segmentation. A second serious difficulty associated with speech perception is the fact that the units of speech often lack clear boundaries. The exact degree of the severity of this problem depends upon what units one selects as the basic units of speech perception, but the problem is present no matter what the unit might be.

Consider the two kinds of units we will be primarily concerned with in this paper -- phonemes and words. First, with respect to phonemes, it is clear that they are not like letters in printed words. Successive phonemes overlap with each other in time and it is not only difficult but we believe also counter-productive to impose boundaries between them. At the best, segmentation throws away some of the information about the identity of the phoneme. At the worst, it can mislead other processes, since it is so prone to error. What we seek, then, is a model which identifies the sequence of units in a speech stream without requiring or imposing artificial boundaries between the segments.

With respect to words, there are some aspects of the articulation of an utterance that give some clues to the locations of word boundaries. But these clues are not always reliable, and in fact the end of one word often runs into the beginning of the next, with the phoneme or phonemes at the boundary belonging to more than one word at a time. Thus, again, segmentation prior to word recognition will necessarily be an errorful process, and in some cases (unless overlapping segmentations are allowed), would perforce discard some part of one of the words. Again, it appears that we need a model which identifies the sequence of words in a speech stream without the deleterious interference of a prior segmentation process.

Immediate exploitation of constraint. Marslen-Wilson (1980) has reviewed a number of findings which illustrate what we will call the **immediacy principle**. The principle states that information can be used by the human perceiver of speech very shortly after if becomes available. Thus, for example, when the acoustic stream unambiguously determines the identity of a word, by ruling out all words but one, it appears that this information is exploited right away in speeding the detection of a target phonemes.

While it was arrived at on the basis of psychological experimentation, the immediacy principle must be taken as an important computational challenge. For example, models such as HEARSAY (Carnegie-Mellon University, 1977; Erman & Lesser, 1980) which take whole utterances into a buffer and process them in an opportunistic fashion, do not exhibit this sort of behavior, and part of the appeal of Marslen-Wilson's COHORT model lies in the fact that it assumes that the speech perception process can operate in real time.

Robust mechanisms of lexical access. Thompson (1984) has recently pointed out that lexical access mechanisms must be capable of functioning properly with noisy input information anywhere in the word, including at the very beginning. In addition, he points out, they must be capable of exhibiting what he calls the "right context effect"; information coming after a particular segment must be able to constrain the identity of what came before it, since segments early in a word are prone to ambiguity and distortion as much as those coming later.

Here is seems that we have something of a dilemma from a computational point of view. On the one hand, according to Marslen-Wilson's immediacy principle, we want a mechanism which exploits constraint as soon as it becomes available; on the other hand, we need a mechanism that can keep its options open so that it can exploit right as well as left contextual influences. We will see below that these two computational constraints can be jointly satisfied through interactive activation.

Psychological Phenomena

Trading relations. Even if we restrict our attention to the direct acoustic information associated with a phoneme, there are generally a number of distinct cues to each phoneme's identity. Sometimes some of these cues may be absent altogether, and in any case they can vary considerably in strength. It appears that human listeners exploit whatever cues are available (Cole, 1973). It is therefore possible to trade-off one cue against another and still obtain reliable perception of the phoneme. These *trading relations* have been documented in a large number of different studies (e.g., Denes, 1955; Derr & Massaro. 1980; Massaro & Cohen, 1977; Repp, 1981, 1983).

Categorical Perception. Even though we are flexible in the cues we use for phoneme identification, as indicated by the trading relations experiments, we are nevertheless rather absolute, or *categorical*, about the result of the perceptual process. This rigidity is exemplified in two coupled phenomena. First, there tends to be a rather sharp boundary between phonemes along any given dimension of variation, particularly for stop consonants. That is, for any given dimension that can be varied systematically we may find that one region of the continuum gives rise to the perception of one phoneme, while another gives rise to the perception of another phoneme; and that the zone of transition between the two percepts is very sharp. To be sure, there is a transitional zone, but relative to the "within-category" part of the continuum, the "between-category" part is rather small, and outside of it there is little uncertainty about the phoneme's identity. The second, intrinsically coupled aspect of the phenomenon of categorical perception is the fact that subjects appear to be very poor at discriminating between different speech sounds that are identified as belonging to the same category; discrimination is much better for sounds that straddle the category boundary.

Thus, there is something of a paradox. The existence of trading relations makes it appear that we are exquisitely sensitive to the strength of each of the different cues that go into the determination of phoneme identity, and can play one cue off against another. On the other hand, it appears that we hear in terms of phonetic categories, rather than in terms of the actual acoustic signals that give rise to these perceptions. We will see that the interactive activation mechanism provides a simple computational mechanism which explains this paradoxical behavior.

Perceptual constancy with varying cues. The ability to trade one cue off against another is one tribute to the flexibility of the human perceptual system. Another is the ability to perceive the same phoneme, even when the cues are altered as a function of context. For example, one cue to the distinction between /b/ and /w/ at the beginning of a word is the rise time of the amplitude envelope as the lips open in uttering the /b/ or /w/. Short rise times are interpreted as /b/'s and longer ones as /w/'s. Just how long a rise time can be before it is interpreted as a /w/ varies as a function of the duration of the following vowel. When the following vowel is very short, an intermediate rise-time will be interpreted as a /w/, whereas if the following vowel is long, the same rise-time will be interpreted as /b/ (Miller, 1981). Thus, it would appear that the perceptual system retunes its detectors for particular phonemes in a way that is sensitive to the parameters of the context.

This particular finding of Miller's is by no means an isolated phenomenon. While there are some cues to phoneme identity which are relatively invariant, analysis of natural speech indicates that many cues shift as a function of phonetic context. Even cues such as the cue known as formant locus (c.f., Delattre, Liberman, & Cooper, 1955) which is often taken as a possible invariant cue to place of articulation of stop consonants, turn out to vary systematically with the following vowel (Ohman, 1966; Fant, 1973; Kewley-Port, 1982); and perceptual studies show that human subjects have no difficulty correctly perceiving the invariant phoneme, in spite of the variability of these perceptual cues. This comes about, we believe, via the exploitation of the lawful variability in the speech wave that we

mentioned in the section on computational constraints.

Lexical effects on phoneme identification. There are a large number of different studies that show that there are lexical influences on the identification -- and perceptual experience -- of phonemes. For example, Ganong (1980) has described a phenomenon he calls the "Lexical Effect", in which the identity of an ambiguous phonetic segment (e.g., a /g/ or a /k/) is determined by the lexical status of the utterance in which it is embedded. Thus, this ambiguous segment will tend to be heard as /g/ when followed by /ift/ but as /k/ when followed by /is/, because "gift" and "kiss" are words but "kift" and "giss" are not. Similarly, Samuel (1981) has shown that the "phonemic restoration effect", the tendency to fill in perceptually phonemes which have been excised from utterances and replaced by noise, depends on the lexical status of the string from which the phoneme is excised.

But lexical effects are not always obtained. They depend on a variety of factors, and a number of researchers have been tempted to conclude that lexical effects on phoneme identification are not real, or are "just guessing", based on their absence in certain circumstances. One of our goals for the TRACE model is to show that the model can account, not only for the cases in which lexical effects do occur, but for the cases in which they do not occur as well.

Effects of phonotactic rules. Massaro and Cohen (1983) have recently reported that subjects tend to perceive ambiguous phonemes in accordance with a system of rules which specifies the legal phoneme sequences in English. In word-initial position, the sequences /sl/ and /tr/ are phonotactically acceptable in English but /sr/ and /tl/ are not. Massaro and Cohen (1983) found that a segment which is ambiguous between /r/ and /l/ will tend to be perceived as /l/ when preceeded by /s/, but as /r/ when preceeded by /t/. Such findings offer seem to cry out for the assumption that the phonotactic rules of a language are explicitly represented in the perceptual mechanisms. In TRACE, however, we find that such phenomena emerge from the interactions of units standing for particular words that embody

these phonotactic constraints; no explicit rules are required.

Lexical basis of word segmentation. As we have already pointed out, the acoustic cues to the boundaries between words are unreliable. It should be no surprise, then, that the segmentation of sequences of phonemes into words is based, at least in part, on the lexical status of different possible sequences. Little detailed experimental work has been done on this phenomenon, but there are several clear-cut illustrative examples that can be cited of this sort of influence. Consider the sequences /Sikant/ and /sikant/. The first of these will be heard as the two words "She can't", while the second will be heard as the single word "secant". Doubtless higher-level linguistic factors above the word level come into play here, but a major contributor to this effect, we have supposed, is the fact that "shecant" is not a single word, but "secant" is. As we shall see later, there are a number of recent phenomena in speech perception which fit together with this analysis.

The model we shall describe in the rest of this paper is designed to meet the computational challenges we have described, and to account for the psychological phenomena. The rest of the paper has three main parts. First, we describe the basic assumptions of the TRACE model. Second, we present simulations of TRACE I, a version of the model which accepts real speech, and show how it takes steps toward meeting some of the computational challenges at the phoneme level. Third, we present simulations using TRACE II, a version of the model which accepts "simulated" input that retains some of the properties of real speech, and show how the model can accommodate the computational challenges at the word level and can account for lexical effects on phoneme identification and the lexical basis of word segmentation.

The TRACE Model

Overview

The TRACE model consists primarily of a very large number of nodes, organized into three levels, the feature, letter, and word levels. Orthogonal to these three levels, the nodes are organized into groups associated with successive moments, or time slices.

Input to the model, in the form of patterns of activation to be applied to the nodes at the feature level, is presented sequentially to the feature-level nodes in successive slices. Thus, at any instant, input is arriving only at the nodes in one slice at the feature level. However, all the nodes are continually involved in processing, and processing of the input arriving at one set of slices is just beginning as the input is moved along to the next set of slices.

The entire network of nodes is called "the Trace", because the pattern of activation left by a spoken input is a trace of the analysis of the input at each of the three processing levels. This trace is unlike many traces, though, in that it is dynamic, since it consists of activations of processing elements, and these processing elements continue to interact as time goes on.

Processing takes place through the excitatory and inhibitory interactions of the units in the Trace. Nodes on different levels that are mutually consistent have mutually excitatory connections, while nodes on the same level that are inconsistent have mutually inhibitory connections. All connections are bi-directional. Unlike the interactive activation model of visual word recognition (McClelland and Rumelhart, 1981), there are *no between level inhibitory connections*. This simplifies the structure of the model, while at the same time increasing its ability to perform the tasks we want it to accomplish, as we shall see in later sections.

Nodes and Connections that make up the Trace

The following sections describe the detailed assumptions about the nodes and the connections that make up the Trace at each of the three levels.

Feature level nodes and connections. At the feature level, the nodes correspond to detectors for acoustic features of the speech stream at particular moments in time. The nodes are organized into sets standing for different values along the same featural dimension. There is a node for each of eight different values on each of a large number of dimensions in each slice of the Trace.

Figure 1 illustrates the detectors for each the eight different values along one dimension, in a single time-slice. Nodes for features on the same dimension are mutually inhibitory; this mutual inhibition is represented by the lines connecting all of the nodes on the dimension.

Phoneme level and feature-phoneme connections. There is one node for every phoneme in every time-slice of the Trace. Each phoneme node in a particular slice receives excitatory input from feature nodes in a range of slices, extending both forward and backward from the slice in which the phoneme node is located. The size of the phoneme's "window" on the feature level varies as a function of the intrinsic duration of the phoneme and as a function of the tendency of the feature to be spread out in time. However, the strength of the input is greater from those feature slices centered under a particular phoneme node than it is for those at greater distances; the falloff is linear, reaching zero at different distances from the center.

There are inhibitory connections between nodes at the phoneme level. In essence, nodes inhibit each other to the extent that the units they stand for overlap in time. Thus, nodes for phonemes centered in the same slice will inhibit each other more strongly than those centered in adjacent slices; again the fall-off is linear and reaches zero at a distance that depends on the length of the phoneme.

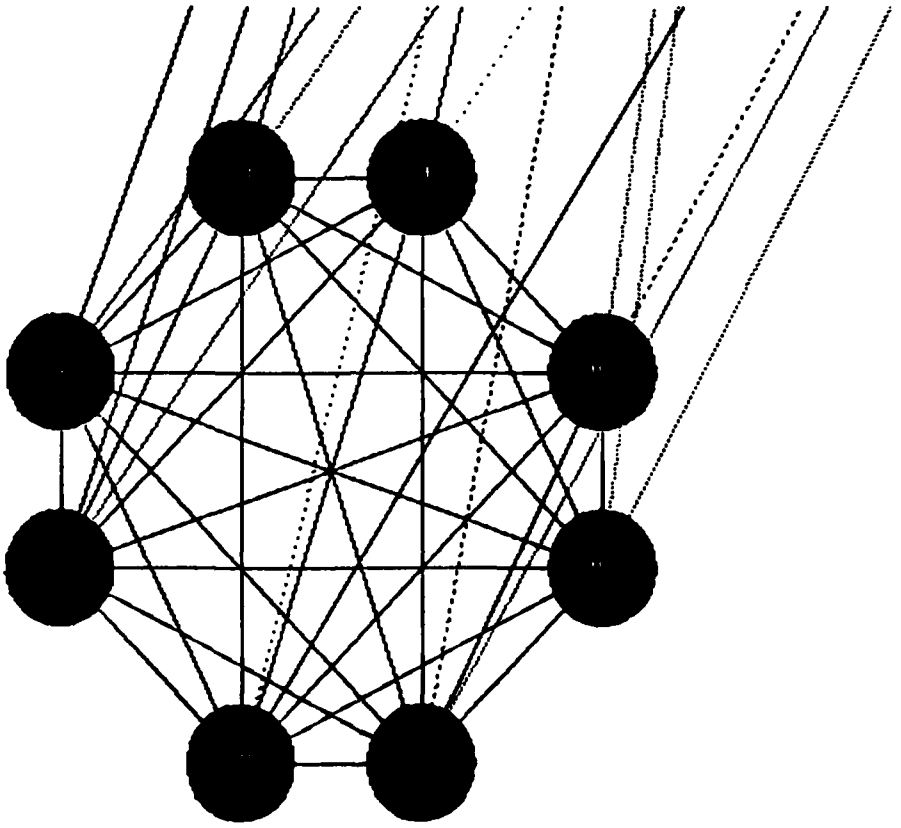


Figure 1. The detectors for each of the eight different values along one feature dimension, in a single time-slice. Note that there is one group of eight nodes for each dimension, in each time slice of the Trace.

These assumptions are illustrated in Figures 2 and 3. Figure 2 shows two feature dimensions and two phoneme nodes from a particular time slice, and their excitatory and inhibitory connections. The bi-directional excitatory connections between feature and phoneme levels are indicated by the lines between the feature and phoneme nodes. Note that there are no between-level inhibitory connections. Note also that only some of the feature values on a given dimension excite any particular phoneme node, and that the strengths of these excitatory connections can differ, so that some values on a dimension activate a particular phoneme more strongly than others. These variations in strength are suggested by the different qualities of some of the lines between the feature and phoneme levels. In Figure 3, we see a series of these displays from an oblique perspective. There is one copy of the nodes and connections shown in the first panel for each of several successive time slices. Some of the connections between units in adjacent slices have been added. Though it is difficult to see this without very close inspection of the figure, one of the two illustrated phoneme nodes (for /a/) has excitatory connections with feature nodes in two adjacent slices on either side, while the other (for /p/) has excitatory connections with feature nodes in only the one immediately adjacent slice on either side, illustrating the differences in "widths" of the different phonemes. In addition, the /a/ nodes have bi-directional inhibitory connections at the phoneme level that extend further forward and backward than the /p/ nodes.

Word nodes and word-phoneme connections. There is a node for every word in every time slice. Each of these nodes represents a different hypothesis about a word identity and starting-location in the trace. For example, the node for the word /gift/ in slice 4 represents the hypothesis that the input contains the word gift starting in slice 4. More exactly, it represents the hypothesis that the input contains the word gift with its first phoneme centered in slice 4.

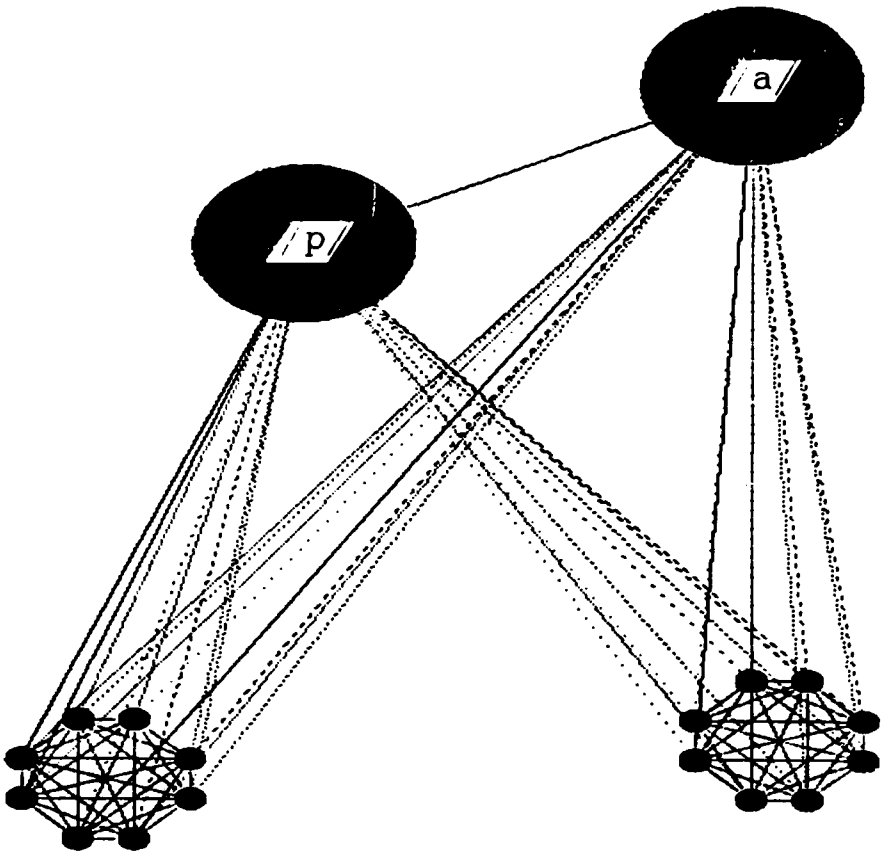


Figure 2. Two feature dimensions and two phoneme nodes from a particular time slice, and their excitatory and inhibitory connections.

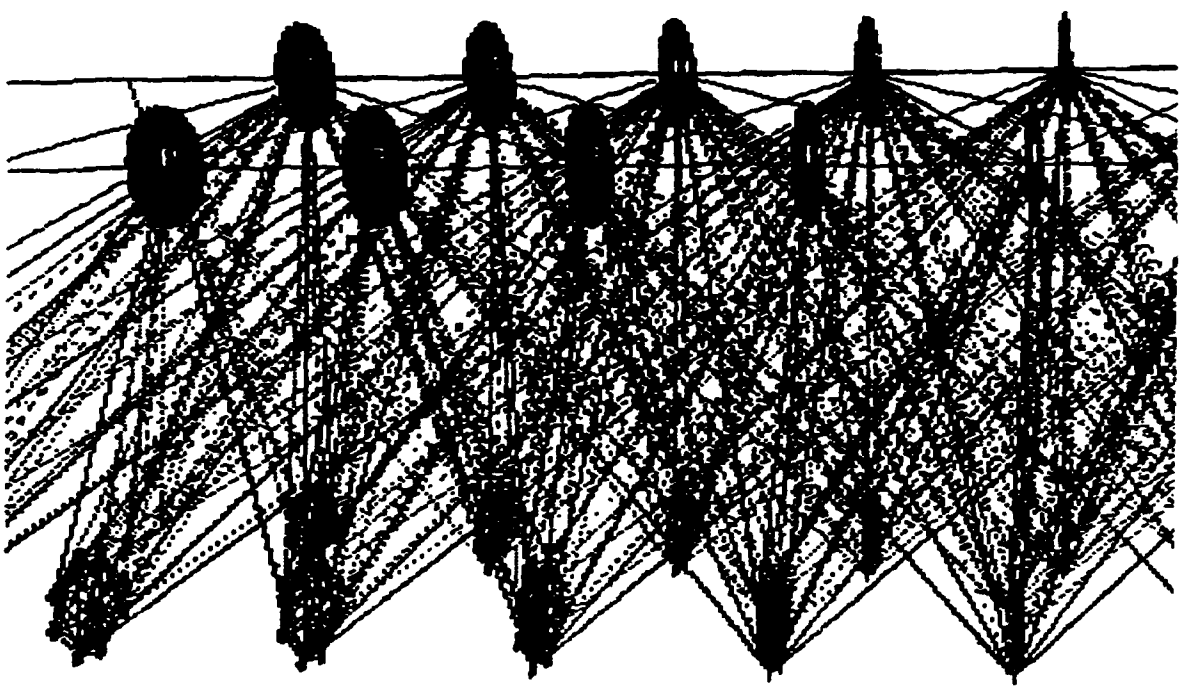


Figure 3. Copies of the same set of nodes for several successive time slices, illustrating the fact that each phoneme level node interacts with feature level nodes in several adjacent time slices.

Word nodes receive excitation from the nodes for the phonemes they contain in a series of overlapping windows. Thus, the node for GIFT in time slice 6 will receive excitation from /g/ in several adjacent slices centered on slice 6, from /i/ in several adjacent slices centered several slices further along, etc. As with the feature-phoneme connections, these connections are strongest at the center of the window and fall of linearly on either side, reaching zero at a point that depends on the duration of the phoneme.

The inhibitory connections at the word level are similar to those at the phoneme level. Again, the strength of the inhibition between two word nodes depends on the number of time slices in which they overlap. Thus, nodes representing alternative interpretations of the same stretch of phoneme nodes are strongly competitive, but interpretations of non-overlapping sequences of phoneme nodes do not compete at all. Figure 4 illustrates a very small number of word nodes and phoneme nodes from a single slice. In Figure 5, a number of slices have been concatenated, and excitatory and inhibitory connections between units in adjacent slices have been added.

Processing An Utterance

Before processing of an utterance begins, the activations of all of the nodes are set at their resting values. At the start of processing, the input to the first slice of feature nodes is applied. Node activations are then updated, ending the first time cycle. On the second time cycle, the input to the second slice of feature nodes in applied, and excitatory and inhibitory inputs to each node resulting from the pattern of activation left at the end of the first time slice are computed. It is important to realize that the interactive active activation process is occurring throughout the Trace on each time slice, even though the external bottom-up input is only coming in to the feature nodes in a single slice at a time.

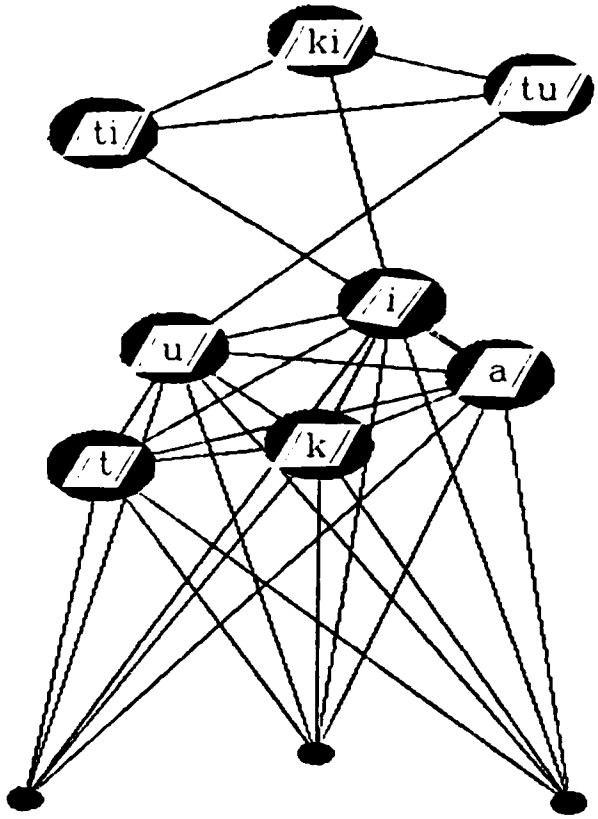


Figure 4. A few phoneme and word nodes within a single slice of the Trace.

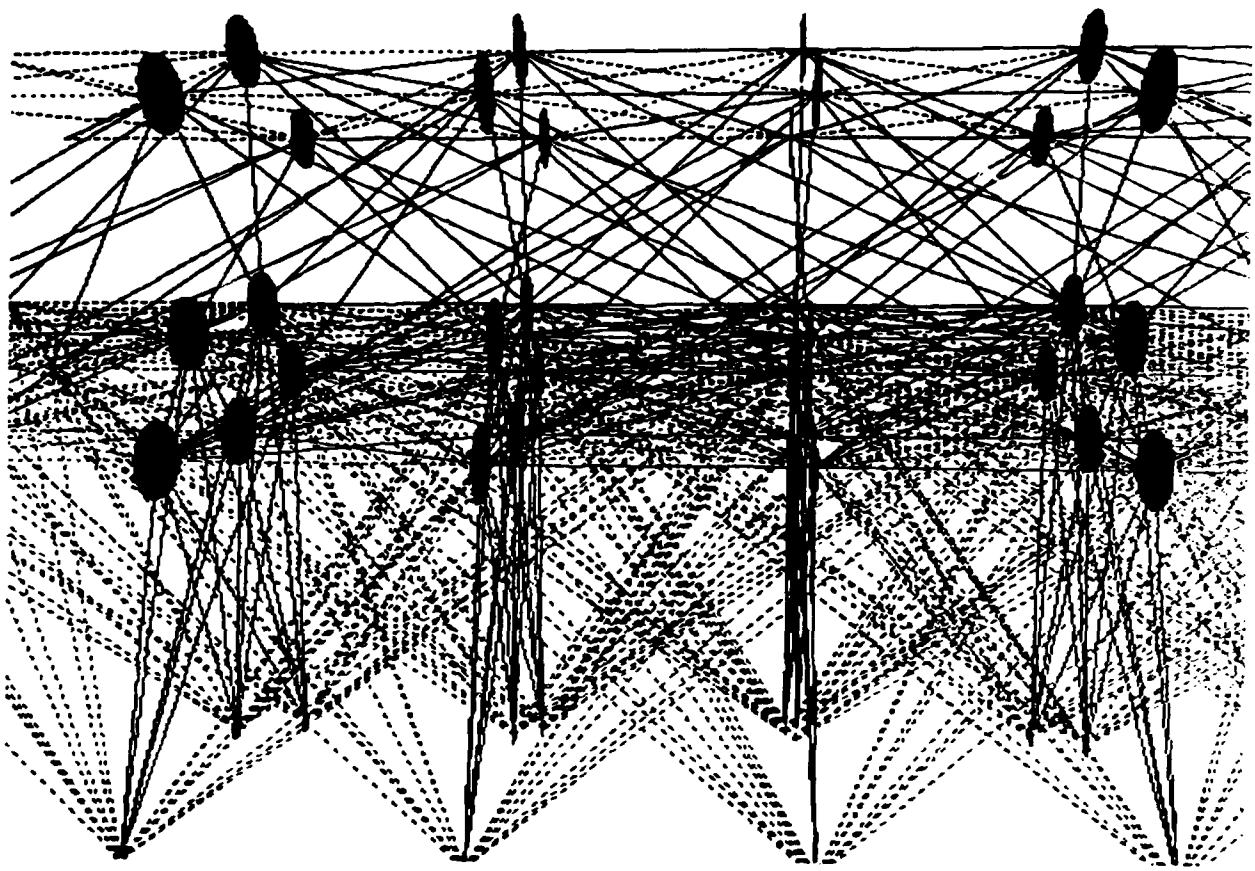


Figure 5. Several replicas of the network from Figure 4 viewed obliquely, one for each of several slices, with connections between and within slices.

The input specification of a single utterance generally spans between 50 and 100 slices. Processing interactions can continue beyond the end of the input, however. Once the input runs out, there are simply no new input specifications applied to the Trace; the continuing interactions are based on what has already been presented. This interaction process is assumed to continue indefinitely, though for practical purposes it is always terminated after some number of time cycles has elapsed.

Details of processing dynamics. The interactive activation process in TRACE follows the dynamic assumptions laid out in McClelland and Rumelhart (1981). Each node has a resting activation value arbitrarily set at 0, a maximum activation value arbitrarily set at 1.0, and a minimum activation set at -.2. On every time-cycle of processing, all the weighted excitatory and inhibitory signals impinging upon a node are added together. The signal from one node to another is just the extent to which its activation exceeds 0; if its activation is less than 0, the signal is 0. The weights are determined by the strengths of the connections between the nodes in question, as spelled out above, and by global level-specific constants.

After the net input to each node has been determined based on the prior activations of the nodes, the activations of the nodes are all updated for the next processing cycle. The new value of the activation of the node is a function of its net input from other nodes and its previous activation value. The exact function used (see McClelland and Rumelhart, 1981) keeps node activations bounded between their maximum and minimum values. Given a constant input, the activation of a node will stabilize at a point between its maximum and minimum that depends on the strength and sign (excitatory or inhibitory) of the input. With a net input of 0, the activation of the node will gradually return to its resting level.

Output Assumptions

Activations of nodes in the Trace rise and fall as the input sweeps across the feature level. At any time, a decision can be made based on the pattern of activation as it stands at that moment. The decision mechanism can, we assume, be directed to consider the set of nodes located within a small window of adjacent slices within any level. The nodes in this set then constitute the set of response alternatives, designated by the identity of the unit for which the node stands (note that with several adjacent slices included in the set, several nodes in the alternative set may correspond to the same overt response). The decision mechanism can be asked to make a response either a) at a criterial time during processing, or b) when a unit in the alternative set reaches a criterial strength relative to the activation of other alternative units. Once a decision has been made to make a response, one of the alternatives is chosen from the members of the set. The probability of choosing a particular alternative i is then given by the Luce (1959) choice rule:

$$p(R_i) = \frac{S_i}{\sum_j S_j}$$

when j indexes the members of the alternative set, and

$$S_i = e^{ka_i}$$

The transformation ensures that all activations are positive and gives great weight to stronger activations, and the Luce rule ensures that the sum of all of the response probabilities adds up to 1.0. Substantially the same assumptions were used by McClelland and Rumelhart (1981).

Parameters and other details.

The particular sets of feature dimensions used, the sizes of the feature and letter-level windows used, and the values of the connections from the feature level to the phoneme level were established separately for the two versions of the model. The details of these choices are described separately for each version of the model below. Once these basic decisions had been made, there remained a set of global, level-specific constants, determining the rate of decay of node activations at each level, the strengths of the excitatory connections between nodes at each pair of adjacent levels, and the strengths of the inhibitory connections between nodes on the same levels. These global constants are the tunable parameters of the model; slightly different values are used in TRACE I and TRACE II, primarily because the operate on different time scales. Within each version of the model separately, all of the reported simulations were carried out with a single set of values for these parameters, with one exception that will be noted. In general, the behavior of the model was extremely robust over variations in the parameters; adjustments of particular parameters influenced the size of particular effects, and sometimes their timing, but did not determine whether a particular effect would occur or not.

Sources of TRACE

The TRACE model grows out of a number of different sources. One of these can be traced back to the Hearsay Speech understanding system (Lowerre & Reddy, 1980); Hearsay represented a radical departure in parallel, interactive processing, and lead to the development of Rumelhart's (1977) interactive model of Reading, and from there to the interactive activation model of word recognition (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1981). We hope the present paper pays back some of the debt this work in reading owes to the study of speech perception.

Another source of inspiration comes from work on neural modeling, particularly by Grossberg (1978, 1980) and Anderson (1977; Anderson, Silverstein, Ritz, and Jones, 1977). Grossberg's (1978) model of memory played an important role in our early thinking about building a model of speech perception, and he deserves particular credit for his analysis of the behavior of competitive inhibitory networks (Grossberg, 1973). While our model is not tied to any particular neural structures, it clearly derives some of its inspiration from the apparent parallel architecture of the brain, and the work of Anderson and Grossberg has pioneered the application of such neural architectures to cognitive and perceptual problems.

We also owe a debt to what we might call the computational connectionists -- those who have applied highly-parallel, connectionist architectures to computational problems such as vision. In particular, Hinton's (1981) connectionist implementation of a variable mapping between retinocentric and object-centered *coordinates* was the basis for our use of activation dependent connection strength modulation. Similar ideas have been proposed by Feldman and Ballard (1982).

A third source of inspiration comes from psycholinguistic work on human speech perception. Particularly important here was the empirical work of Marslen-Wilson (e.g., Marslen-Wilson and Welsh, 1978; Marslen-Wilson, 1980) in demonstrating the interactive nature of speech perception. While we feel we have improved upon it, we nevertheless owe a great debt to the Cohort Model for breaking the ground and laying out a psychological model of the perceptual process. Other sources of inspiration in the psychology of speech perception include the logogen model of Morton (1969) and the information integration models of Massaro and Oden (1980) We have also drawn some insights from the model of Nusbaum and Slowiaczek (1982); their model is closely related to our own, though to date they have focussed on slightly different aspects of speech perception than those we focus on here.

Trace I: Steps Toward a Processor for Real Speech

The principle challenge that has been set to TRACE I is to identify the phonemes contained in CV syllables, and in so doing to illustrate how interactive activation mechanisms can be used to meet two of the computational challenges facing mechanisms of speech perception. These are a) to identify the sequence of phonemes from a continuous speech stream without segment boundaries and b) to exploit the lawful variability in the speech wave.

Details of TRACE I

Input to the model. Spoken inputs to the model consisted of consonant-vowel syllables consisting of one of the three voiced stops /b/, /d/, and /g/, followed by one of the three vowels, /i/ (as in "beet", /a/ as in "father", and /u/ as in "boot". Fifty different tokens of each syllable were recorded. A preprocessor extracted fifteen different parameters from the acoustic signal at successive 5 msec intervals. The parameters and their interpretations are listed in Table 1.

Feature detectors. The values of each parameter were used to generate specifications of excitatory inputs to be applied to the feature level of the model. TRACE I treated each parameter as a separate dimension. Thus, the model contained complete sets of eight detectors for each parameter, for each 5-msec segment of the input. Distributions of parameter values were compiled over all of the different stimuli, and these aggregate distributions were used to find the maximum and minimum values for each parameter. The range between the maximum and minimum was then divided into eight value-ranges, or octiles. One detector on each dimension was assigned to respond maximally to the midpoint value of each octile, with a fall-off to half maximum response at the midpoint of the adjacent octile. In this manner, discrete parameter values as extracted by the preprocessor produced a distribution of activation over two or three detectors in each dimension, for each 5 msec time slice of the speech signal.

Table 1

The fifteen different dimensions used in coding the real speech inputs to TRACE I

power	log rms power
pitch	fundamental frequency of source
alpha	total LPC error; distinguishes speech and silence
edr	error dynamic range; voiced from voiceless
abrupt	change in rms power over time; stops/vowel transitions
diffuse	second moment of LPC spectrum
acute	gross concentration of energy in spectrum
consonantal	smoothed euclidean spectral difference; detects stop bursts and consonantal transitions
spectrum[3]	total energy in each section of the spectrum divided equally in three on a log scale
formant[3]	the values of the first three formants
loci[2]	the hypothetical formant onsets, back-extrapolated from the actual formant onsets (at voicing) to the stop release;

Phoneme nodes and feature-to-phoneme connection strengths. At the phoneme level, TRACE I contained complete sets of detectors for phonemes spaced at 15-msec intervals -- thus, there were three times as many slices at the feature level as at the phoneme level.

The connection strengths used in TRACE I for the feature -- phoneme connections were determined as follows. Half of the stimuli were used to "train" the feature -- phoneme connection values. The other half were used to test the model, as we shall describe below. For the training phase, 25 tokens of each phoneme were used to select connection strengths for the feature to phoneme connections using a modified version of the Perceptron Convergence Procedure of Rosenblatt (1962). The procedure involves presenting either the consonant or the vowel portion of a speech sound and its correct identity, and adjusting feature -- phoneme connection strengths whenever the existing set of connection strengths leads the model to provide a stronger net bottom-up input to any phoneme other than the correct phoneme. For the present experiment, there was one training phase in which the connections were trained for the three consonants, regardless of the vowel context, and a second training phase in which the connections were trained for the vowels, regardless of consonant context. In this way, sets of weights were derived for both the stops and the vowels.

We do not wish to claim either sufficiency or psychological validity for the particular set of parameters used, or the details of the perceptron convergence procedure as a method of settling on connection strengths. We have, however, done our best to use parameters which have often been proposed as candidates for invariant features of phonemes, and to use a learning procedure for finding connection strengths that is known to be able to find a perfect set if one exists. However, it should be noted that our procedure finds a single set of connection strengths which best characterize a particular phoneme throughout its entire temporal extent. An alternative would be to specify the phoneme in terms of the shape of this evolving pattern. However, it should be noted that some of the parameters (locus, abruptness, consonantality) are abstractions from the shape of the evolving pattern, so some of

this information is captured in any case.

Simulations with TRACE I

An example simulation run. Our first simulation simply illustrates that TRACE I behaves reasonably well in identifying the phonemes in one of the CV syllables that was not used during the training phase of the experiment. The pattern of activation produced by one of these CV syllables is shown in Figure 6. The Figure shows activations of phoneme nodes in successive 15 msec slices. Activation is indicated by the height of the symbol, with the symbol itself indicating the identity of the phoneme the node stands for. In this case, nodes for /b/ are most strongly activated by the consonant portion and /a/ by the vowel portion -- this is as it should be, since the syllable was in fact /ba/, as indicated by the lettering along the bottom. Note that the model has already identified both the consonant and the vowel, even though the input is just beginning to move into the vowel. This is because the feature values characteristic of /a/ are blended with those of the /b/, even from the very beginning. Since the /a/ detectors look for input over very wide windows, the phoneme level is able to anticipate the vowel quite early on.

Representation without segmentation. The simulation run just described illustrates how TRACE I is able to produce activations of nodes corresponding to the correct phonemes, without relying on any prior segmentation of the input into separate phonemes. Indeed, the simulation illustrates how the model actually uses the information about two adjacent phonemes that is blended together or, as Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967) put it, "encoded" in each portion of the speech stream. Even while the model is still receiving input from that part of the speech stream that we would identify with the consonant, it is nevertheless beginning to activate nodes representing its best guess as to the identity of the following vowel. Thus the model makes use of the information that is distributed across overlapping portions of the speech stream, rather than

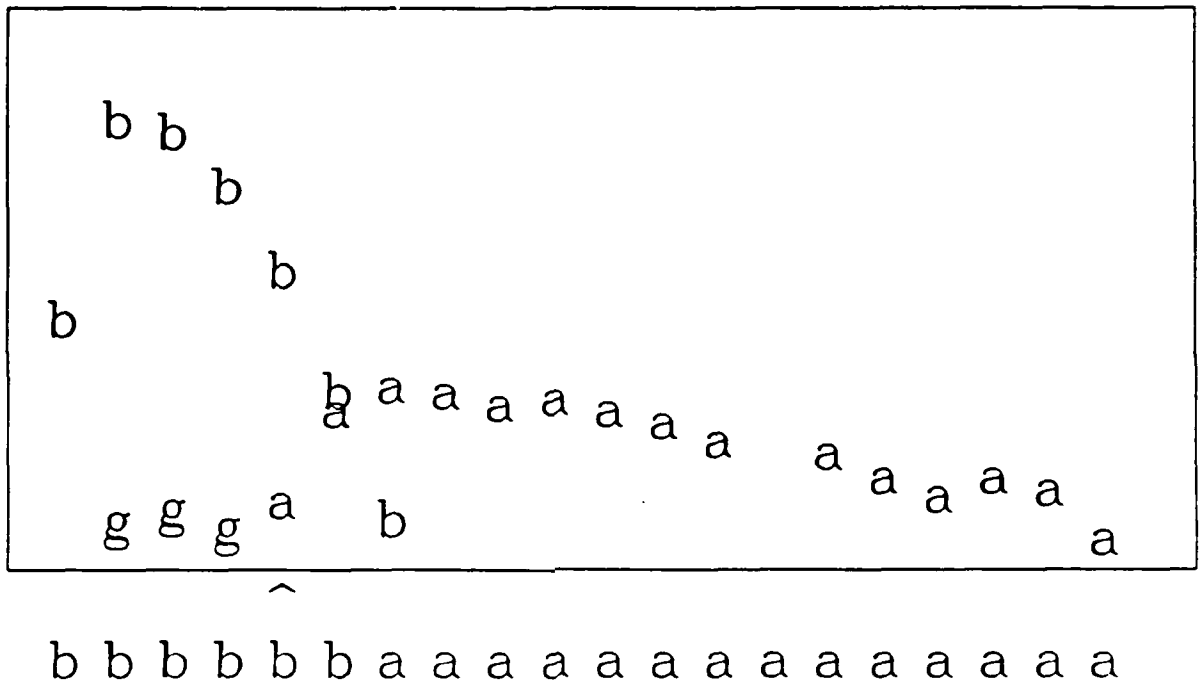


Figure 6. Activations of phoneme nodes in different slices of the Trace, part-way through processing a token of the syllable /ba/. A rough characterization of the locations of the phonemes in the input is given along the ordinate; the "^" symbol indicates which part of the input is currently being directed into the corresponding slice of the Trace. Portions of the stream to the left of the "^" have already been presented; portions of the stream to the right remain to be presented. The top and bottom edges of the figure correspond to activation values of .9 and 0.0. respectively.

simply throwing large parts of this information away, as mechanisms which segment the stream before phonetic analysis are apt to do.

Exploiting Lawful Variability. While TRACE I nicely illustrates the ability to identify phonemes without prior segmentation, it does not do quite as well at phoneme identification as we might desire. In this section, we describe how TRACE I can be enhanced to do this, in a way that is basically consistent with the principles of interactive activation. Before doing this, though, we present the results of a simulation experiment using the model as already described, to give a baseline for assessing the performance of the enhanced version of the model.

In this simulation experiment, we presented all 25 test tokens of each of the 9 CV syllables to the model, and simply asked, for each stimulus, which consonant received the strongest activation. We then compared this to the correct answer to see how often the most strongly activated phoneme corresponded to the correct answer. The activation of each stop was simply the sum of the activations of all tokens of that stop, between time slices 0 and 10. The frequency with which the largest sum was associated with the correct consonant is shown in Table 2, for each CV syllable. Overall, the largest sum was associated with the correct answer 79% of the time.

A number of reasons could be cited for the less-than-perfect performance of the model. Perhaps we are extracting the wrong parameters from the spoken input, perhaps our procedures for extracting them could be improved, and perhaps our use of the phoneme as basic unit is the source of the problem (Fujimura & Lovins, 1978). We believe, though, that a large part of the problem is that we have not, up to this point, been exploiting the lawful variability in the speech wave. The parameter values of the different phonemes are very different when they occur in different contexts. While some cues to phoneme identification are more invariant than others, experimental evidence suggests there are some cues which people use which are not invariant and there is evidence that people exploit the law-

Table 2

Accuracy of Phoneme Identification in 25 Trials
With Fixed and Variable Weights
(percentage correct)

[ba]		[bI]		[bu]	
variable	fixed	variable	fixed	variable	fixed
25 (100)	17 (68)	14 (56)	13 (52)	25 (100)	20 (80)
[da]		[dI]		[du]	
variable	fixed	variable	fixed	variable	fixed
24 (96)	21 (84)	21 (84)	20 (80)	23 (92)	21 (84)
[ga]		[gI]		[gu]	
variable	fixed	variable	fixed	variable	fixed
22 (88)	21 (84)	24 (96)	22 (88)	25 (100)	24 (96)

Overall percentage correct for variable weight condition: 90

Overall percentage correct for fixed weight condition: 79

ful variability in many of these cues (e.g., Alfonso, 1981; Massaro & Oden, 1980a, 1980b; Miller & Eimas, 1977; Oden & Massaro, 1978; Derr & Massaro, 1980; Repp, Liberman, Eccardt, & Pesetsky, 1978). Any mechanism which relies on an invariant characterization of each phoneme, regardless of context, is inherently limited because it fails to exploit this lawful variability.

One possible solution would be to replace the phoneme unit with another unit, such as the diphone or demisyllable (Fujimura & Lovins, 1978). We do not wish to take a firm stand against such proposals, but we do wish to point out that none of the units which have been proposed are immune to variation as a function of their context of occurrence. We have, therefore, sought to find a way of exploiting lawful variability, sticking for the moment with the phoneme as the basic unit, not because of any particular preference for the phoneme, but precisely because phonemes so clearly do vary as a function of their context. The method we propose for exploiting this variability with phonemes could be applied to other units as well.

The essence of the situation is this. The parameters of any speech sound, such as a word-initial stop consonant, vary as a function of of the context in which the sound occurs. To exploit this variability in an interactive activation model, we can is allow the strengths of the connections from the feature level to the phoneme level to be modulated as a function of the context as well. That way, to the extent that we think that the vowel in a syllable is /a/, we can invoke the mapping from feature activations to phoneme activations that is appropriate for /a/.

This idea of using the activation of one unit to determine the strength of the connection between two other units has been proposed both by Hinton (1981) and by Feldman and Ballard (1982). The use of this mechanism here was inspired by Hinton's use of the same device to implement a variable mapping between retinocentric and object-centered feature detectors in a model of visual information processing.

To see how well this approach can work, we developed an enhanced version of TRACE I which modulates interconnection strengths as a function on context in just this way. As before, we began by generating sets of connection strength values, using the same set of 25 training tokens of each CV syllable that we used before. As before, we also used the Perceptron Convergence Procedure to settle on connection strengths. This time, however, separate sets of connection strengths for the consonants were found for each vowel context, and separate sets of connection strengths for the vowels were found for each consonant context.

The sets of connection strengths used in testing the model were modulated by the activations of phonemes in the context. The more a particular phoneme is active in one part of the Trace, the more strongly the connections into the detectors for other phonemes in adjacent parts of the Trace are weighted toward values appropriate to that particular phoneme.

Details of the connection strength modulation scheme. The quantitative details of the connection strength modulation scheme are given for completeness, but are not necessary to understand the basic idea. For the node for a particular phoneme indexed by i in a particular time-slice indexed by j, at a particular time t during the course of processing, the set of weights for that phoneme node can be designated by the vector $\mathbf{w}_{ij}(t)$, where the elements of the vector are just the individual weights from each feature node to each phoneme node. The formula for $\mathbf{w}_{ij}(t)$ is

$$\mathbf{w}_{ij}(t) = \frac{\sum_k \mathbf{w}_i^k C_{kj}(t)}{\sum_k C_{kj}(t)}$$

where $\mathbf{w}_i^k$ is the vector of weights associated with phoneme i in the context of phoneme k, and where $C_{kj}(t)$ stands for the strength of phoneme k in the time slices adjacent to time-slice j at time t. The

formula for $C_{kj}(t)$ is

$$C_{kj}(t) = e^{\sum_{j\cdot} \omega_{j\cdot} a_{kj\cdot}(t)}$$

where $j\cdot$ ranges over the time slices adjacent to time-slice j and $\omega_{j\cdot}$ weights the contribution of the activation of the node for phoneme k in slice $j\cdot$. The values of $\omega_{j\cdot}$ were selected so that the largest contribution to the contextual strength would come at a distance of + or - 15 slices (75 msec) from slice j, falling off linearly on both sides of this peak.

This enhanced version of TRACE I, with variable connection strengths, was tested with the same 25 test tokens of each of the 9 different CV syllables. As illustrated in Table 2, the model did better with each of the 9 different syllable types, and improved its overall performance to 90% correct.

Generalizing knowledge of predictable variability. This approach to dealing with predictable context effects turns out to have another benefit as well. The knowledge that TRACE has about the interactions between vowel spectra and cues for consonantal place of articulation is sufficiently general that it applies to cases in which a new vowel is presented to the system.

TRACE "knows" about the context effects induced by the vowels /a/, /i/, and /u/, in the sense that the strength of interconnections between specific nodes is modified according to context. The effects that the vowels /a/, /i/, and /u/ have on consonants are, to some extent (particularly for ACUTEness), generalizable to other vowels. That is, a new vowel with ACUTEness similar to one or more of the existing vowels should activate those vowels enough that the coarticulatory effects of the unknown vowel will be compensated for by the weight-modulations brought about by the known vowels.

This prediction was verified by presenting TRACE with the syllable /de/. /e/ is an unknown vowel to TRACE but this sound resembles the vowel /i/ enough that it activates the /i/ node. The effects these two vowels have on the pronunciation of neighboring sounds are similar enough that the /i/ node mimics (to a lesser extent) the effect which a node for /e/ *would* have (were it known to the system) and retunes the feature-phoneme mappings for the /d/ node, thereby greatly improving the model's performance in identifying the /d/, as illustrated in Figure 7. The Figure shows what happens when the syllable /de/ is presented with the weight-modulations induced by the vowels turned on (panel a) or off (panel b). The better performance in Figure 7a demonstrates clearly that the coarticulatory information--even for novel vowels--provides a very important source of information about the stop.

The level of accuracy we have been able to achieve with the help of connection strength modulation is comparable to that which has been achieved by other machine based phoneme identification systems (Kopec, 1984). Obviously, though, we have a long way to go before we can claim to have achieved absolute computational sufficiency at the level of phoneme identification. Human observers have no difficulty identifying the stimuli we tested our model on, and it goes without saying that their capabilities extend considerably beyond the limited domain in which TRACE I has been successful. We hope, however, that we have illustrated the fruitfulness of connection-strength modulation as a way of extending the computational sufficiency of interactive activation mechanisms. The use of activation-dependent connection strength modulation makes it possible to exploit the lawful relations between the auditory patterns of speech units and the contexts in which they occur in a graded, and flexible fashion.

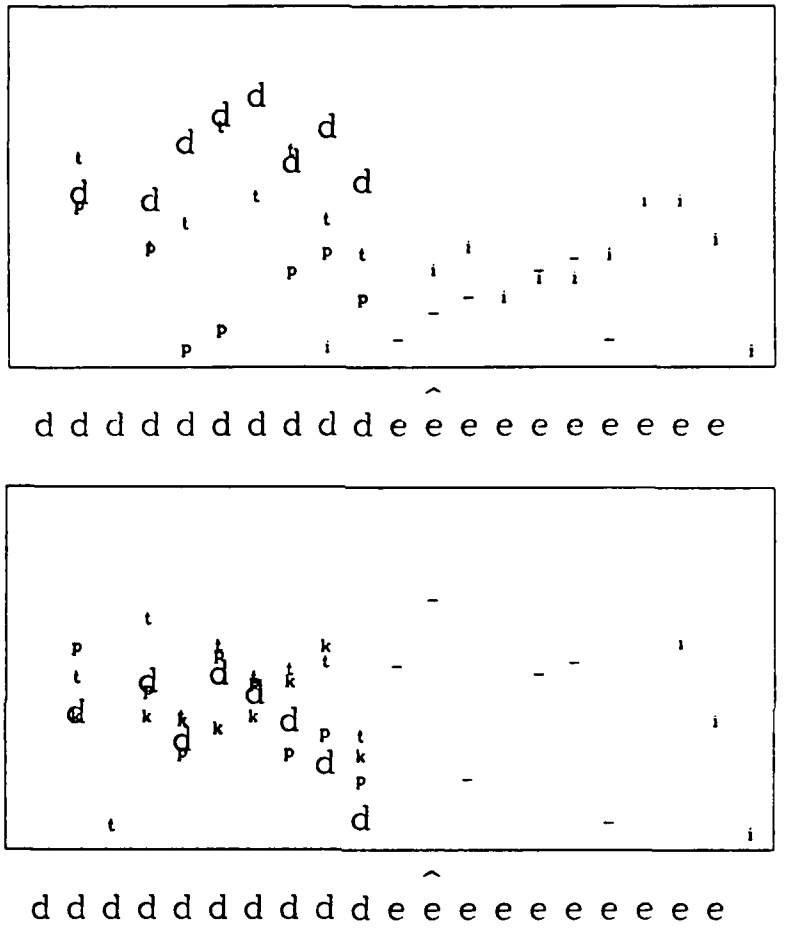


Figure 7. Activation levels of phonemes nodes resulting from the input /de/, with variable connection strengths enabled, in a), and disabled, in b). Nodes competing with /d/ have been displayed in smaller print.

TRACE II: The Phoneme and Word Levels, and their Interaction.

Overview

In this section, we use "simulated" speech inputs to a slightly different version of TRACE to illustrate the model's ability to meet several computational criteria and to account for several psychological phenomena. First, we consider two phenomena at the phoneme level, namely trading relations among different cues to phoneme identity, and categorical perception. Then, we consider interactions between the phoneme and word levels, focusing on the lexical effect on phoneme identification (together with the conditions under which this effect is not obtained), and the apparent influence of phonotactic constraints on phoneme identification. After that, we consider TRACE II in light of two computational criteria for word recognition, namely the immediacy principle and robustness in the face of imprecise input. In the last set of simulations with TRACE II, we illustrate how the model meets the additional computational criterion of identifying words imbedded in continuous streams of phonemes, and we show how the word identification mechanisms in TRACE can be used to account for some data on how humans determine where one word ends and the next one begins.

Before we turn to these simulations, we describe the detailed characteristics of this version of the model.

Details of TRACE II.

Simulated speech input. The input to TRACE II is not real speech, but a series of feature-node activation specifications generated by a simple computer program. A simplified set of only seven dimensions was used in TRACE II to encode the feature level inputs. Though the dimensions were adapted (with extensions) from classical work in phonology (Jakobson, Fant, and Halle, 1952), absolutely nothing is claimed about their psychological validity or the validity of the values assigned to

phonemes on these dimensions, or about the validity of the details of the way in which the feature patterns were generated. These are intentional simplifications of the real structure of speech, in much the same way that the font used by McClelland and Rumelhart (1981) in their model of visual word recognition was an intentional simplification of the real structure of print.

The feature patterns do have the following properties that are important for our analysis.

1) The feature patterns used for phonemes which are similar perceptually were similar to each other. We do not claim to have captured the full details of phoneme similarity, but the feature sets do have the property that the feature pattern for one phoneme will tend to excite nodes for similar phonemes more than the nodes for less similar phonemes.

2) The feature patterns were constructed in such a way that it was possible to create feature patterns that would activate two different phonemes to an equal extent by averaging the values of the two phonemes on one or more dimensions. In this way, it was a simple matter to make up ambiguous inputs, half-way between two phonemes, or to construct continua varying between two phonemes on one or more dimensions.

3) The patterns for successive phonemes overlapped, as they do in real speech. The feature pattern corresponding to a particular phoneme extended over several time slices of the input (specifically, 11). The strength of the pattern grew to a peak at the 6th slice and fell off again. Peaks of successive phonemes were separated by 6 slices.

4) There were no cues in the speech stream to word boundaries -- the feature specification for the last phoneme of one word overlapped with the first phoneme of the next in just the same way feature specifications of adjacent phonemes overlap within words. However, entire streams of speech, be they individual syllables, words, or strings of words, are preceeded and followed by silence, and we assume that silences provide cues that are exploited when they are available. Therefore, we used silences in the input to delimit the beginnings and endings of entire streams. In TRACE II, silence was not simply the absence of any input; rather, it was a pattern of feature values, just like the phonemes. Thus, a value on each of the seven dimensions was associated with silence. These values were actually beyond the range of values which occurred in the actual phonemes, so that the features of silence would not activate any of the phoneme nodes.

To detect the presence of silence, a ninth node was added to each feature continuum to serve as a silence feature detector. The model also includes nodes that serve as detectors for "silence units" both at the phoneme and at the word level. Thus, there are "phoneme silence" and "word silence" nodes in TRACE II. The input to the phoneme silence units comes from a window on feature silence

nodes; the inputs to the word silence units comes from a window on the phoneme silence units. At both the phoneme and the word levels, a silence was as long as a single regular phoneme; of course, there could be sequences of silences, but in all the simulations the inputs to TRACE II began and ended with single silences.

In general, things were kept simple in TRACE II by using homogeneous parameters wherever possible. Thus, as already noted, the feature specifications of all phonemes were spread out over the same number of time slices, effectively giving all phonemes the same duration. The strength of the total excitation coming into a particular phoneme node from the feature nodes was normalized to the same value for all phonemes, thus making each phoneme equally excitable by its own canonical pattern. Other simplifying assumptions were adopted as well. For example, there were no differences in connections or resting levels for words of different frequency. (See McClelland and Rumelhart, 1981, for a treatment of the effects of word frequency in visual word recognition).

The repertoire of phonemes used in TRACE II consisted of the stop consonants /b/,/d/,/g/,/p/,/t/, and /k/; the fricatives /s/ and /S/ (sh); the liquids /r/ and /l/; and the vowels /a/ (as in father), /u/ (as in pool) /i/ (as in tree) and /^/ (as in but); /^/ was also used for reduced, unstressed vowels such as the second vowel in "target".

The set of words known to the model consisted of 211 words meeting all of the following constraints: a) The word consisted only of the phonemes given above; b) It was not an inflection of some other word that could be made by adding "-ed", "-s" or "-ing"; c) the word and its "-ed" "-s" and "-ing" inflections occurred with a frequency of 20 or more per million in the Kucera & Francis (1967) word count. It is not claimed that the model's lexicon is an exhaustive list of words meeting this criterion, since the Kucera-Francis word list is not coded phonetically, but it is reasonably close to this.

Time-slices in TRACE II should be thought of as being considerably coarser than those used in TRACE I, where parameter values from real speech inputs were extracted at 5 msec intervals. In TRACE II, we think of each slice as corresponding roughly to 25 msec, so that the time from the peak of one phoneme to the peak of the next (6 slices) is roughly 150 msec.

As in TRACE I, the time-slices at the phoneme and word levels were spaced more widely than at the feature level. That is, there were three feature slices to every phoneme and word slice; phoneme and word nodes, therefore, are spaced at approximately 75 msec intervals. In speaking of the location in the Trace of a node, we refer to its position in feature-slice units, to avoid confusion. Thus, when we refer to the node for /b/ in slice 12 we mean the node for /b/ that is centered over feature slice 12.

It should be noted that the model took no account of syllable stress or any other factor on vowel duration or indeed on the feature specifications of the phonemes; the only contextual influence on the successive phoneme specifications was simple overlapping. Where they overlapped, the feature node activations specified by two adjacent phonemes were simply added together. This was considered adequate for present purposes, which do not concern the computational sufficiency of the phoneme level.

Perceptual Phenomena at the Phoneme Level

In this section we consider trading relations and categorical perception. As we have seen, these two phenomena, taken separately, seem to cast the mechanisms of speech perception in quite different lights. Here we see how both emerge naturally from the interactive activation mechanisms of TRACE.

For these simulations, the model was stripped down to the essential minimum necessary, so that the basic mechanisms producing trading relations and categorical perception could be brought to the

fore. The word level was eliminated altogether, and there were only four phonemes in the model: /a/, /b/, /d/, and silence (/-/). From these four phonemes, inputs and percepts of the form /-ba-/ and /-da-/ could be constructed. The following additional constraints were imposed on the feature specifications of each of the phonemes: 1) the /a/ and /-/ had no overlap with either /b/ or /d/, so that neither /a/ not /-/ would bias the activations of the /b/ and /d/ phoneme units where they overlapped with the consonant. 2) /b/ and /d/ were identical on five of the seven dimensions, and differed only on the remaining two dimensions.

The two dimensions which differentiated /b/ and /d/ were a) Acuteness and b) Burst Amplitude; /b/ is less acute that /d/ and has a lower-amplitude burst. The pattern of excitatory input to the Acuteness and Burst Amplitude units on the feature level produced by the canonical /b/ and the canonical /d/ are illustrated in Figure 8.

Trading relations. TRACE quite naturally tends to produce trading relations between features, since it relies on the weighted sum of the excitatory inputs to determine how strongly the input will activate a particular phoneme node. All else being equal, the phoneme node receiving the largest sum bottom-up excitation will be more strongly activated than any other, and will therefore be the most likely response when a choice must be made between one phoneme and another. Since the net bottom up input is just the sum of all of the inputs, no one input is necessarily decisive in this regard.

Generally, experiments demonstrating trading relations between two or more cues manipulate each of the cues over a number of values ranging between a value more typical of one of two phonemes and a value more typical of the other. The results of a classic experiment of this type, by Denes (1955) are shown in Figure 9. The main finding is that the probability of choosing one of the two alternatives varies with the manipulation of each of the two cues, such that there are cases in which a cue that favors one of the two phonemes to a moderate degree will give rise to the perception

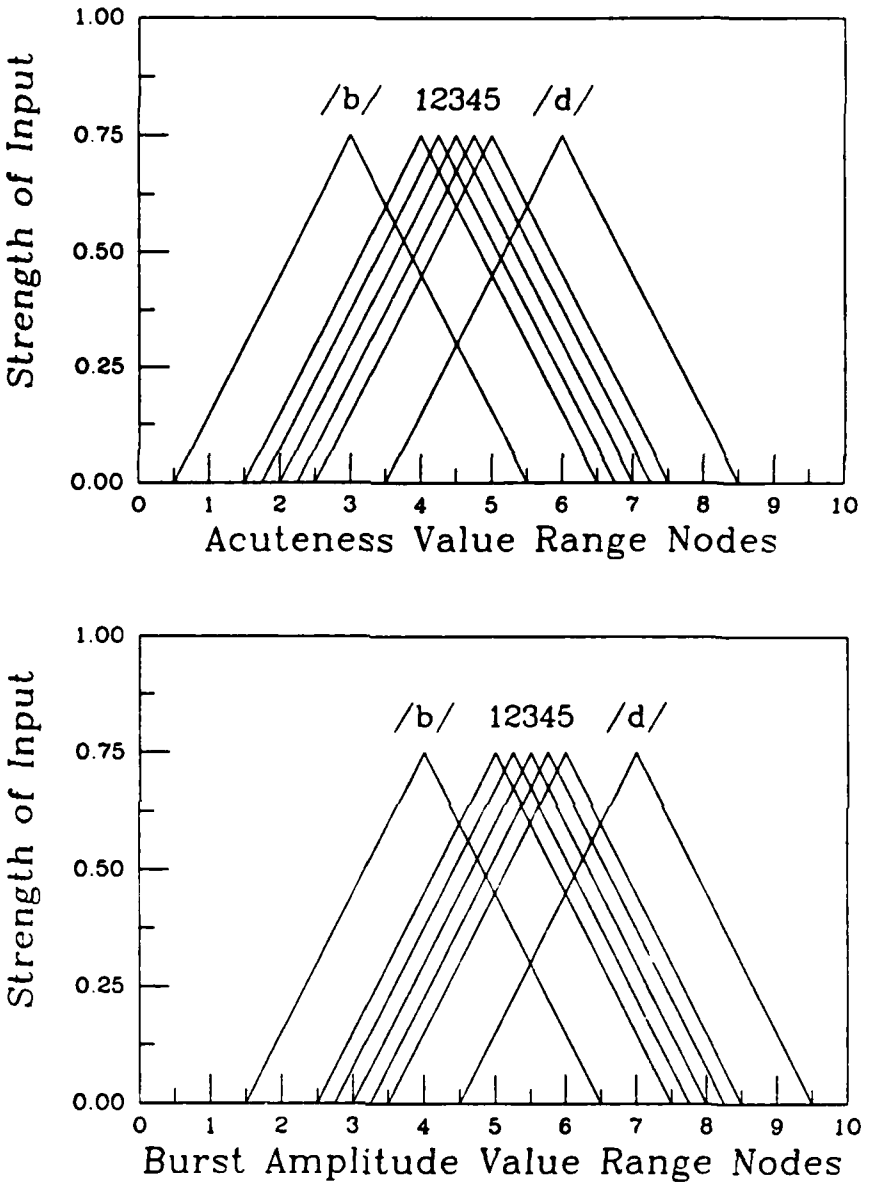


Figure 8. Canonical feature level input for /b/ and /d/, on the two dimensions that distinguish them, and the patterns used for the five intermediate values. Along the abscissa of each dimension the nine nodes for the nine different value ranges of the dimension are arrayed. The curves labeled /b/ and /d/ indicate the relative strength of the excitatory input to each of these nodes, for /b/ and for /d/ separately. The canonical curves also indicate the strengths of the feature to phoneme connections for /b/ and /d/; that is, the canonical input pattern for each phoneme exactly matches the strengths of the corresponding feature-phoneme connections. Numbered curves on each dimension show the feature patterns used in the trading relations simulation.

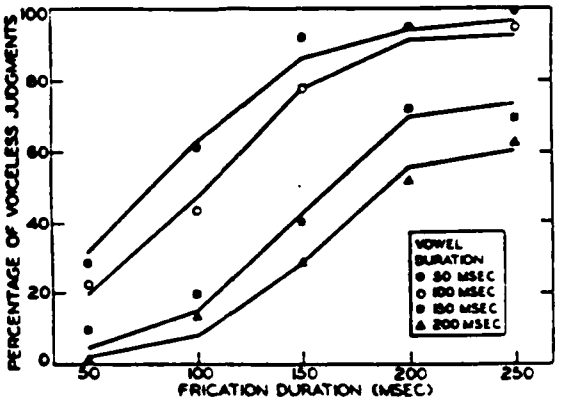


Figure 9. Results of an experiment demonstrating the trade-off between two cues to the identity of /s/ and /z/. Data from Denes, 1955, reprinted in Derr and Massaro, 1980.

of the other phoneme when paired up with a strong cue that favors the other phoneme. An additional finding is the bowing of the curves; they tend to be approximately linear through the middle of their range, but to level off at both ends, where the values on both dimensions agree in pointing to one alternative or the other.

To see if TRACE would simulate these effects, we generated a set of 25 intermediate phonetic segments made up by pairing each of five different input patterns on the Abruptness dimension with each of five different input patterns on the Burst Amplitude dimension. The different feature patterns used on each dimension are shown in Figure 8, along with the canonical feature patterns for /b/ and /d/ on each of the two dimensions. On the remaining 5 dimensions, the intermediate segments all had the common canonical feature values for /b/ and /d/.

The model was tested with each of the 25 stimuli, preceeded by silence (/-/) and followed by /a-/. In this and all subsequent simulations we will report in this paper, the peak of the initial silence phoneme occurred at time slice 6 in the input, and the peaks of successive phoneme segments occurred at 6 slice intervals. Thus, for these stimuli, the peak on the intermediate phonetic segment occurred at slice 12, the peak of the following vowel occurred at slice 18, and the peak of the final silence occurred at slice 24. For each input presented, the interactive activation process was allowed to continue through a total of 60 time slices, well past the end of the input. The state of the Trace at various points in processing, for the most /b/-like of the 25 stimuli, is shown in Figure 10. In this and subsequent figures, activations of the phoneme nodes located between the peaks of the input specifications of the phonemes (at slices 3, 9, 15, etc) have been deleted from the display for clarity (the activations of these nodes generally get suppressed by the model, since the nodes on the peaks tend to dominate them). At the end of the 60th time slice, we recorded the activation of the nodes for /b/ and /d/ in time slice 12, and the probability of choosing /b/ based on these activations. (It makes no difference to the qualitative appearance of the results if a different decision time is used;

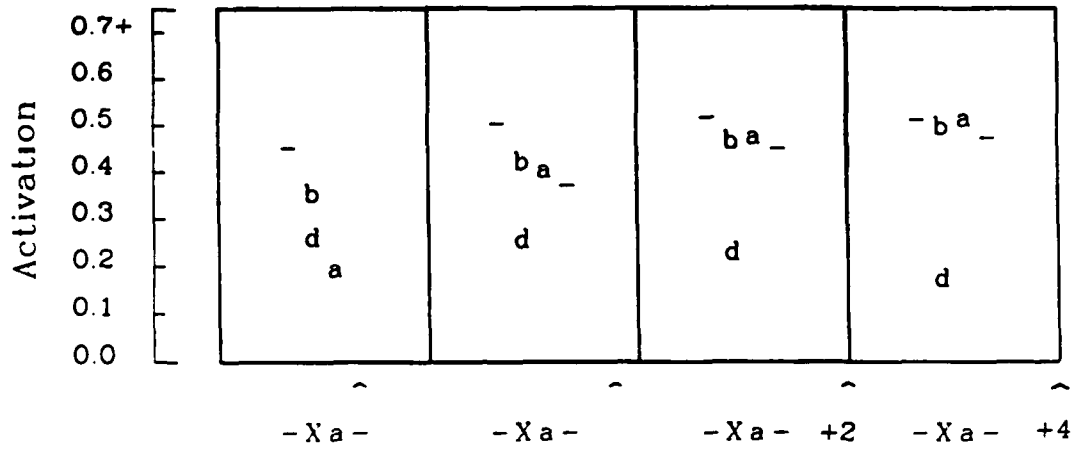


Figure 10. The state of the Trace at various points during and after the presentation of a syllable consisting of the most /b/-like of the 25 intermediate segments used in the trading relations experiment, represented by /X/, preceeded by silence and followed by /a/ then another silence. The peak of the feature specification for each of the segments (including the intermediate segment and silence) is indicated below each panel of the Figure by the location of the corresponding letter.

earlier decision times are associated with smaller differences in relative activation between the /b/ and /d/ phoneme units, and later ones with larger differences, but the general pattern is the same. We picked 60 cycles because the changes had become very gradual by this time, and it seemed unrealistic to imagine human subjects waiting much longer than the 1.5 seconds this roughly represents to make their response.)

Response probabilities were computed using the formulas given earlier for converting activations to response strengths and strengths into probabilities. The resulting response probabilities, for each of the 25 conditions of the experiment, are shown in Figure 11. The pattern of results is quite similar to that obtained in Denes (1955) actual experiment with human subjects: The contribution of each cue is approximately linear and additive in the middle of the range, and but the curves flatten out at the extremes, as in the Denes (1955) experiment; and clearly, the model's behavior exhibits the ability to trade one cue off against another. For example, there are three different combinations of feature values which lead to a probability between acuteness dimension coupled with the most /d/-like value on the Burst amplitude dimension; 2) the neutral value on the Burst amplitude dimension coupled with the most /d/-like value of the Acuteness dimension; and 3) the somewhat /d/-like values on both dimensions.

Categorical perception. In spite of the fact that TRACE is quite flexible in the way it combines information from different features to determine the identity of a phoneme, the model is quite categorical in its overt responses. This is illustrated in two ways: First, the model shows a much sharper transition in its choices of responses as we move from /b/ to /d/ along the Abruptness and Burst Amplitude dimensions than we would expect from the slight changes in the relative excitation of the /b/ and /d/ nodes. Second, the model tends to obliterate differences between different inputs which it identifies as the same phoneme, while sharpening differences between inputs assigned to different categories. We will consider each of these two points in turn, after we describe the stimuli

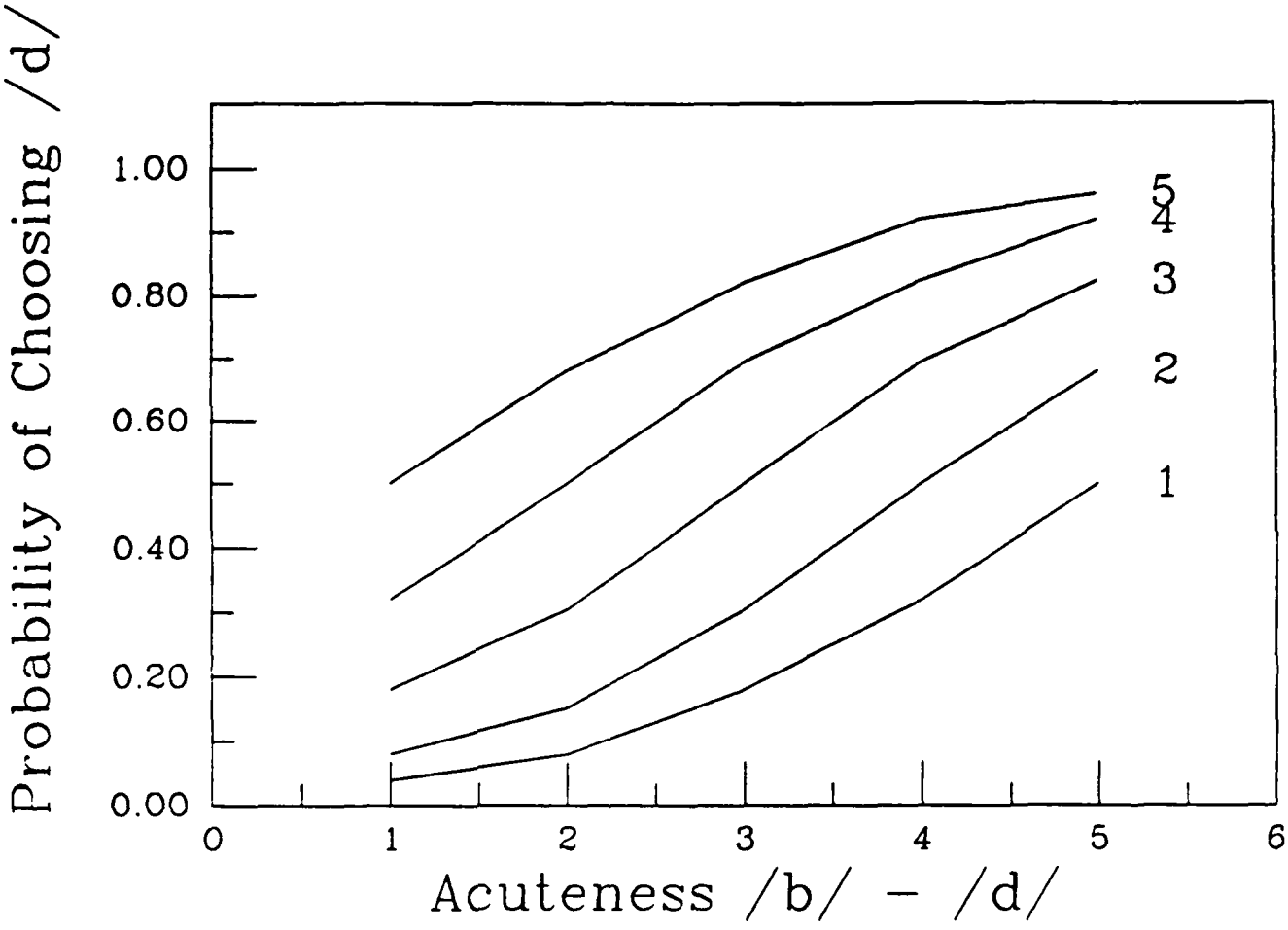


Figure 11. Simulated probability of choosing /d/ at time slice 60, for each of the 25 stimuli used in the trading relations simulation experiment. Numbers next to each curve refer to the intermediate pattern on the Burst Amplitude continuum used in the 5 stimuli contributing to each curve.

used in the simulations.

Eleven different consonant feature patterns were used, embedded in the same /-_a-/ context as in the trading relations simulation. Two of the feature patterns were the canonical /b/ and /d/ feature patterns, followed by the vowel /a/. Between the canonical patterns there were 5 intermediate stimuli, arranged in equal steps between /b/ and /d/ on both of the two continua which differentiated the two consonants. In addition, there were two stimuli on either side of the canonical /b/ and /d/ stimuli. All the stimuli were spaced equal distances apart on the Acuteness and Burst Amplitude dimensions. The locations of the the peak activation values on each of these two continua are shown in Figure 12.

Figure 13 indicates the relative initial bottom-up activation of the /b/ and /d/ phoneme nodes for each of the 11 stimuli used in the simulation. The first thing to note is that the relative bottom-up excitation of the two phoneme nodes differ only slightly. For example, the canonical feature pattern for /b/ sends 75% as much excitation to /d/ as it sends to /b/. The feature pattern two steps toward /d/ from /b/ (stimulus number 5), sends 88% as much activation to /d/ as to /b/.

The Figure also indicates, in the second panel, the resulting activations on the nodes for /b/ and /d/ at the end of 60 cycles of processing. The activation curves exhibit a much steeper transition than the relative bottom-up excitation curves.

There are two reasons why the activation curves are so much sharper than the initial bottom-up excitation functions. The primary reason is *competitive inhibition*. The effect of the competitive inhibition at the phoneme level is to greatly magnify the slight difference in the excitatory inputs to the two phonemes. It is easy to see why this happens. Once one phoneme is slightly more strongly activated than the other, it exerts a stronger inhibitory influence on the other than the other can exert on it. The net result is that "the rich get richer." This general property of competitive inhibition mechanisms was discussed by McClelland and Rumelhart (1981), and has also been noted by Grossberg

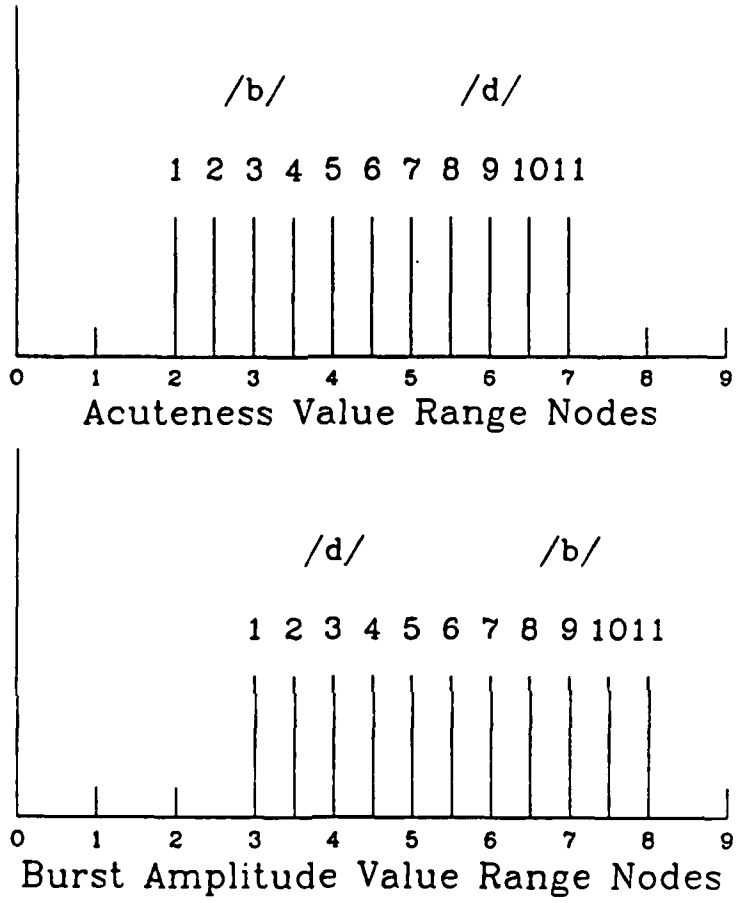


Figure 12. Locations of peak activations along the Acuteness and Burst Amplitude dimensions, for each of the 11 stimuli used in the categorical perception simulation.

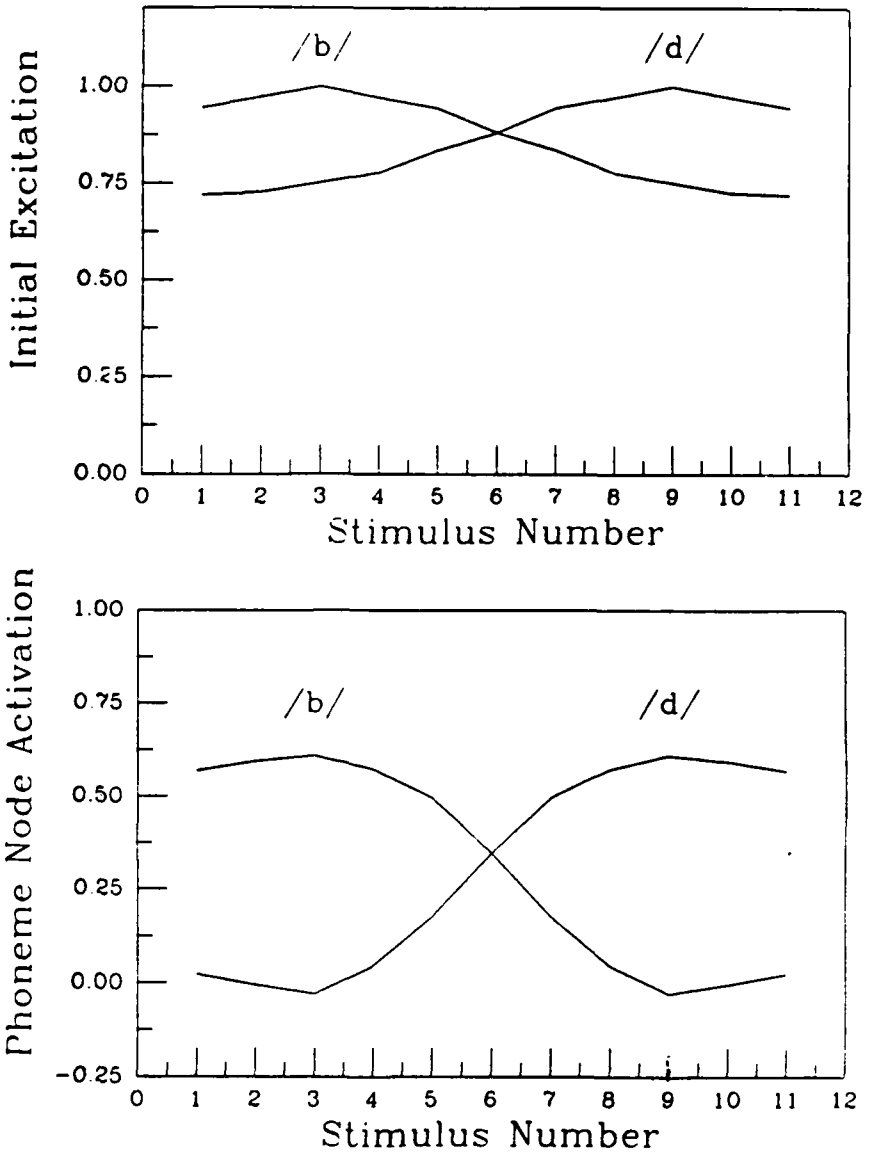


Figure 13. Effects of competition on phoneme activations. The first panel shows relative amounts of bottom up excitatory input to /b/ and /d/ produced by each of the 11 stimuli used in the categorical perception simulation. The second panel shows the activations of nodes for /b/ and /d/ at time cycle 60. Stimuli 3 and 9 correspond to the canonical /b/ and /d/ respectively.

(1978) and Levin (1976); it is also well known as one possible basis of edge enhancement effects in low levels of visual information processing. A second cause of the sharpening of the activation curves is the phoneme-to-feature feedback, which we will consider in detail in a moment.

The identification functions that result from applying the Luce choice rule to the activation values shown in the second panel of Figure 13 are shown in Figure 14 along with the ABX discrimination function, which will be discussed below. The identification functions are even sharper than the activation curves; there is only a 4% chance that the model will chose /d/ instead of /b/ for stimulus 5, for which /d/ receives 88% as much bottom up support as /b/. The increased sharpness is due to the properties of the response strength assumptions. These assumptions essentially implement the notion that the sensitivity of the decision mechanism, in terms of d' for choosing the most strongly activated of two nodes, is a linear function of the difference in activation of the two nodes. When the activations are far enough apart, d' will be sufficient to ensure near-100% correct performance, even though both nodes have greater than 0 activation. Of course, the amount of separation in the activations that is necessary for any given level of performance is a matter of parameters; the relevant parameter here is the scale factor used in the exponential transformation of activations. The value used for this parameter in the present simulations (10) was the same as that used in all other cases where we translate activation into response probability, including the Trading Relations simulation.

Some readers may be puzzled as to why TRACE II exhibits a sharp identification function in the categorical perception experiment, but shows a much more gradual transition between /b/ and /d/ in the trading relations simulation. The reason is simply that finer steps along the Abruptness and Burst Amplitude continua were used in the trading relations simulation. All of the stimuli for the trading relations simulation lie between stimuli 6 and 4 in the categorical perception simulation.

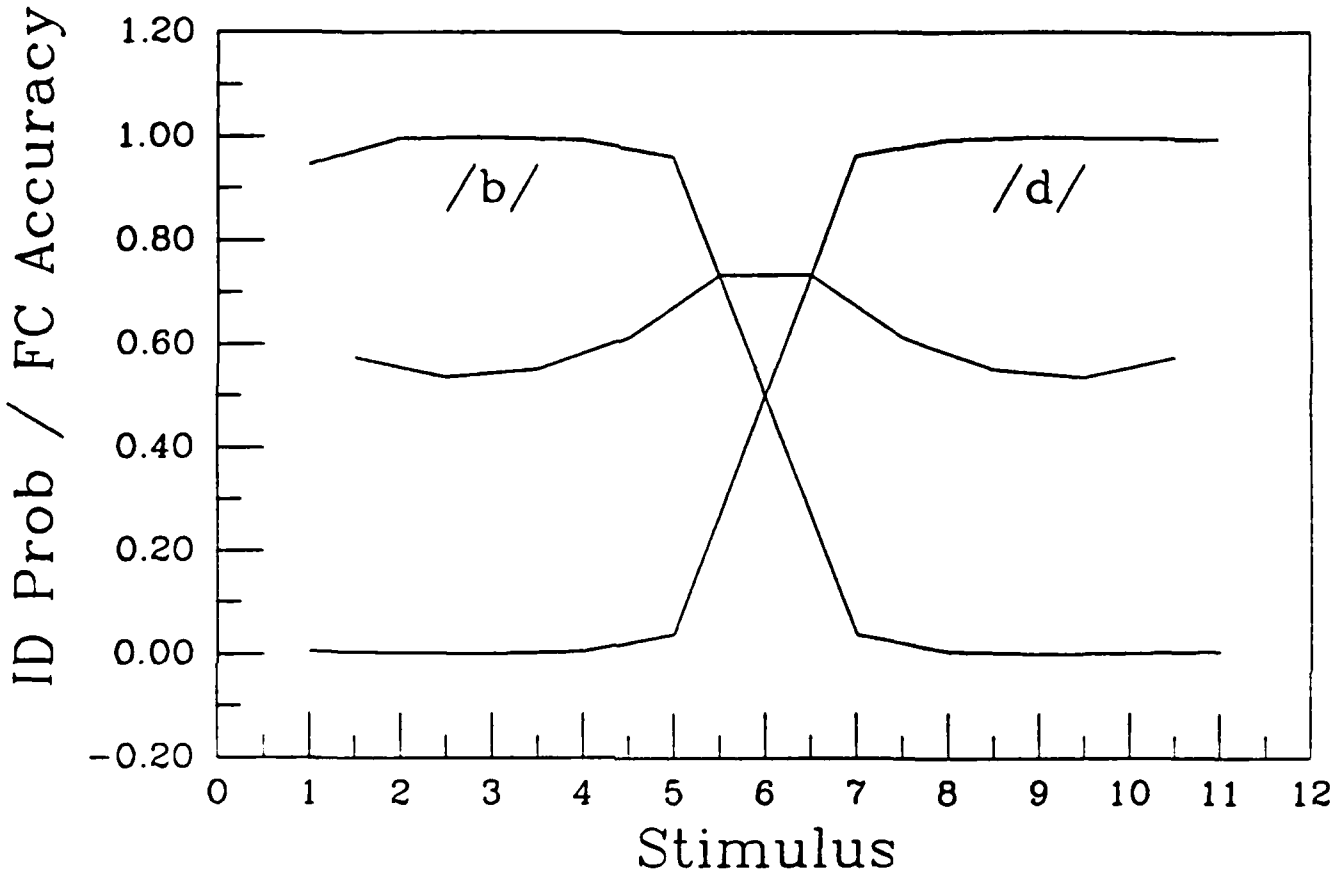


Figure 14. Simulated identification functions and forced-choice accuracy in the ABX task.

This obviously brings out the fact that the apparent steepness of the identification function depends on the grain of the sampling of different points along the continuum between two stimuli, as well as a host of other factors. Whether an empirical or simulated identification function will look steep or not depends on the selection of stimuli by the experimenter or modeler. However, it is worth noting that the steepness of the identification function is independent of the presence of trading relations, at least in the simulation model. That is, if we had used more widely separated steps along the the Acuteness and Burst Amplitude dimension, we would have obtained much steeper identification functions. The additivity of excitatory inputs would still apply, and thus it would still be possible to trade cues off against each other.

The sharpening the model imposes on the identification function, in conjunction with the fact that it can trade one feature off against another, shows how the model, like human perceivers of speech, can be both flexible and decisive at the same time. In fact, the model's decisiveness extends even farther than we have observed thus far; feedback from the phoneme to the feature level tends to cause the model to obliterate the differences between input feature patterns that result in the identification of the same phoneme, thus allowing the model to provide an account, not only for sharp identification functions, but also for the fact that discriminability of speech sounds is far poorer within categories than it is between categories.

Strictly speaking, at least as defined by Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967), true categorical perception is only exhibited when the ability to discriminate different sounds is no better than could be expected based on the assumption that the only basis a listener has for discrimination is the categorical assignment of the stimulus to a particular phonetic category. However, it is conceded that "true" categorical perception in this sense is never in fact observed (Studdert-Kennedy, Liberman, Harris, & Cooper, 1970). While it is true that the discrimination of sounds is much better for sounds which perceivers assign to different categories than for sounds they

assign to the same category, there is also at least a tendency for discrimination to be somewhat better than predicted by the identification function, even between stimuli which are always assigned to the same category. TRACE II produces this kind of approximate categorical perception.

The way it works is this. When a feature pattern comes in, it sends more excitation to some phoneme nodes than others; as they become active, they begin to compete, and one gradually gets the upper hand. This much we have already observed. But as this competition process is going on, there is also feedback from the phoneme level to the feature level. Thus, as a particular phoneme becomes active, it tends to impose its canonical pattern of activation on the feature level. The effect of the feedback becomes particularly strong as time goes on, since the feature input only excites the feature nodes very briefly; the original pattern of activation produced by the phoneme nodes is, therefore, gradually replaced by the canonical pattern imposed by the feedback from the phoneme level. The result is that the pattern of activation remaining at the feature level after 60 cycles of processing has become assimilated to the prototype. In this way, feature patterns for different inputs assigned to the same category are rendered nearly indistinguishable.

An impression of the magnitude of this effect is illustrated in Figure 15, which shows how different two patterns of activation at the feature level are at the end of 60 cycles of processing. The measure of difference is simply $1 - r_{ab}$, where r_{ab} stands for the correlation of the patterns produced by stimuli a and b. Only the two dimensions which actually differ between the canonical /b/ and /d/ are considered in the difference measure. Furthermore, the correlation considers only the feature pattern on the feature nodes in time slice 12, right at the center of the input specification. If all dimensions are considered, the values of the difference measure are reduced, but the overall pattern is the same. Inclusion of feature patterns from surrounding slices likewise makes little difference.

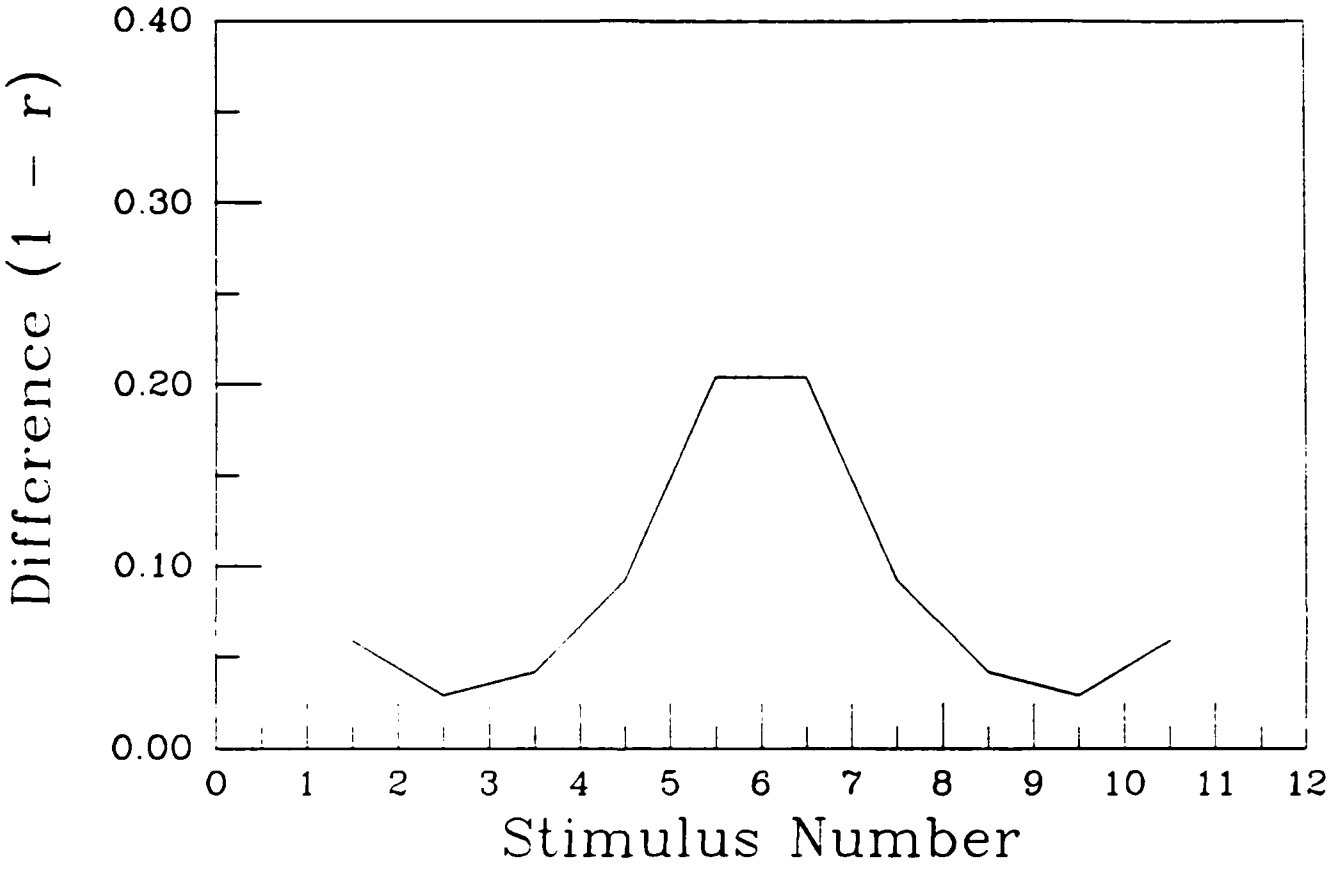


Figure 15. Differences between patterns of activation at the feature level at cycle 60, for pairs of stimuli one step apart along the /b/-/d/ continuum used for producing the identification functions shown previously in Figure 14. The difference measure is the correlation of the two patterns, subtracted from 1.0; thus, if the two patterns correlated perfectly, their difference would be 0.

To relate the difference between two stimuli to probability correct choice performance in the ABX task generally used in categorical perception experiments, we once again use the Luce (1959) choice model. The probability of identifying stimulus x with alternative a in is given by

$$p(R_{x=a}) = \frac{S_{ax}}{S_{ax} + S_{bx}},$$

where S_{ax} is the "strength" of the similarity between a and x. This is given simply by the exponential of the correlation of a and x:

$$S_{ax} = e^{k_r r_{ax}},$$

and similarly for S_{bx}. Here k_r is the parameter that scales the relation between correlations and strengths. These assumptions are consistent with the choice assumptions made for identification responses. The resulting response probabilities, for one choice of the parameter k_r (5) are shown in Figure 14 (the exponentiation parameter k sub r is different than the parameter k used in generating identification probabilities from activations because correlations and activations are not on equivalent scales).

Basically, the figure shows that the effect of feedback is to make the feature patterns for inputs well within each category more similar than those for inputs near the boundary between categories. Differences between stimuli near the prototype of the same phoneme are almost obliterated. When two stimuli straddle the boundary, the feature level patterns are much more distinct. As a result, the probability of correctly discriminating stimuli within a phoneme category is much lower than the probability of discriminating stimuli in different categories.

One interesting property of TRACE's behavior in this simulation is that the discrimination is better for stimuli near the boundary but still within a category than for those close to the center of the category. The primary cause of this difference is the fact that stimuli near the category boundary produce partial activation of both phonemes. Even though one dominates the other sufficiently for near-100% identification, the other is sufficiently activated to produce some feedback of its own, thus preserving at the feature level some of the ambivalence evident higher up.

It should be noted that it would be possible to account for categorical perception in TRACE without invoking feedback from the phoneme level to the feature level. All would we need to do is assume that the feature information that gives rise to phoneme identification is inaccessible for purposes of discriminating phonemes, so that all the perceiver has to go on is the output of the categorical phoneme identification process. However, what is interesting here is that, even allowing access to the feature level, the interactive activation process used in TRACE produces a tendency toward categorical perception. Even if we have access to all levels of processing for the purposes of generating responses, the feedback mechanism tends to impose the results of processing reached at higher levels on lower levels, enhancing differences between items that fall into different categories and obliterating differences between items which the higher levels of processing treat as the same.

We do not wish to suggest the the feedback interpretation of categorical perception is superior to an interpretation which simply claims perceivers have no access to the feature level. The feedback interpretation treats categorical perception as a matter of degree, and allows for all gradations of intermediate cases based on differences in the persistence of the feature activation values, the strength of feedback that might exist for some kinds of features as opposed to others, etc. While the no-access account has the flavor of being more all-or-none, a no-access account can admit differing degrees of access to the feature level for different features as well; indeed, this refuge is generally taken to account for the fact that categorical perception is never absolutely categorical (e.g.,

Studdert-Kennedy, Liberman, Harris, and Cooper, 1970). The fact that vowels show only a weak tendency toward categorical perception can be attributed to greater accessibility, or perhaps persistence, of the feature patterns that give rise to vowel percepts (Fujisaki & Kawashima, 1968). Likewise, in a feedback account, it would be possible to accommodate the fact that vowels and consonants show different tendencies toward categorical perception by assuming that the feature patterns of vowels persist longer in the system than those for consonants; the feedback would therefore exert less of an influence, relative to the initial input, on the eventual pattern of activation at the feature level.

However, the two interpretations do differ in one way that might be testable. The feedback account seems to differ most clearly from a limited feature access account in the case of two stimuli, both away from the center of a category, but still within it. Here, as we have seen, TRACE tends to show greater discrimination than it shows between stimuli squarely in the middle of a category, thus violating the prescription that discrimination be no better than can be accounted for by identification performance, and violating it differentially depending on where within the category the stimuli fall. Such an effect, if it were obtained experimentally, would seem to call for an account that provided for some access to the feature level and gave some reason why patterns nearer the center of a category should be less discriminable than patterns nearer the edge.

There is some evidence bearing on this aspect of TRACE's account of categorical perception. Samuel (1977) has reported ABX discrimination data that show noticeable minima in the discrimination function hear the canonical stimuli within each category on a /d/-/t/ continuum. Indeed, Samuel's account of this effect, though not couched in terms of interactive activation processes, has a great deal of similarity what we see in TRACE; he suggests that near-canonical items are more strongly assimilated to the canonical pattern. Unfortunately the effect we seek is fairly subtle, and so it will be difficult to separate from noise. In Samuel's experiment, the effect is fairly clear-cut in three observers at the end of extensive training, as shown in Figure 16, and even unpracticed subjects tend

to show the effect when stimuli are spaced 6 msec apart along the VOT continuum, as they are in Samuel's "two-step" version of the ABX task.

Lexical Effects on Phoneme Identification

In the previous section, we have seen how competition among phonemes and feedback to the feature level played important roles in making phoneme perception strongly categorical in flavor, while at the same time preserving flexibility in terms of the exact sets of feature values required to produce the perception of a particular phoneme. The present section focuses on how competition at the word level, coupled with feedback to the phoneme level, can produce the lexical effect on phoneme perception.

For these simulations, and for the remainder of the paper, the full set of 15 phoneme nodes and the full lexicon of 211 words were in the model. The feature specifications of the phonemes /b/ and /d/ were a little different, although their overall discriminability was roughly the same as in the previous section. Since we were not explicitly concerned with the feature level in the remainder of the simulations, we eliminated feedback from the phoneme level to the feature level. Two other parameters (rate of decay at the feature level and strength of phoneme-level inhibition) were adjusted to compensate for the effects of this simplification.

You can tell a phoneme by the company that it keeps.[1] In this section, we describe a simple simulation of the basic Lexical effect on phoneme identification, reported by Ganong (1980). For the first simulation, the input to the model consisted of a feature specification which activated /b/ and /p/ equally, followed by (and partially overlapping with) the feature specifications for /I/, then /ˆ/, then /g/. Figure 17 shows phoneme and word level activations at several points in the unfolding of this

1. This title is adapted from the title of a talk by David E. Rumelhart on related phenomena in letter perception. These findings are described in Rumelhart and McClelland (1982). We thank Dave for his permission to adapt the title.

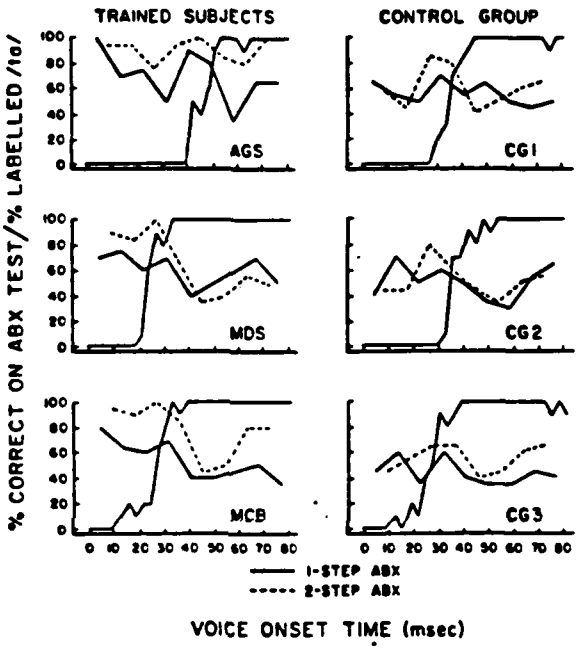


Figure 16. Identification and one and two-step ABX discrimination data from three practiced and three naive subjects from Samuel (1977). One step A and B stimuli were 3-msec apart on the VOT continuum; two-step stimuli were 6 msec apart.

input specification.

The figure illustrates the gradual build-up of activation of the two interpretations of the first phoneme, followed by gradual build-ups in activation for subsequent phonemes. As these processes unfold, they begin to produce word level activations. It is difficult to resolve any word-level activations in the first few frames, however, since in these frames, the information at the letter level simply has not evolved to the point where it provides enough constraint to select any one particular word. In this case, it is only after the /g/ has come in that the model has information telling it whether the input is closer to "plug", "plus" or "blood". After that point, as illustrated in panel d, "plug" wins the competition at the word level, and through feedback support to /p/, causes /p/ to dominate /b/ at the phoneme level. The model, then, provides an explicit account for the way in which lexical information can influence phoneme identification.

Two things about the lexical effect observed in this case are worthy of note. First, the effect is rather small. Second, it does not emerge until well after the ambiguous segment itself has come and gone. There is a slight advantage of /p/ over /b/ in frames 2 and 3 of the figure. In these cases, however, the advantage is not due to the specific information that this item is the word /plug/ -- the model can have no way of knowing this at these points in processing. The slight advantage for /p/ at these early points is due to the fact that there are more words beginning with /pl/ than /bl/ in the model's lexicon, and in particular, there are more beginning with /pl˙/ than /bl˙/. So, when the input is /?l˙d/, with the ? standing for the ambiguous /b/-/p/ segment, the model must actually overcome this slight /p/-ward bias. Eventually, it does so.

To show when the information about the identity of the final phoneme comes in and the extent of its effect, Figure 18 shows the temporal course of buildup of the strength of the /p/ response based on activations of the phoneme nodes in slice 12 for two cases in which the initial segment is ambigu-

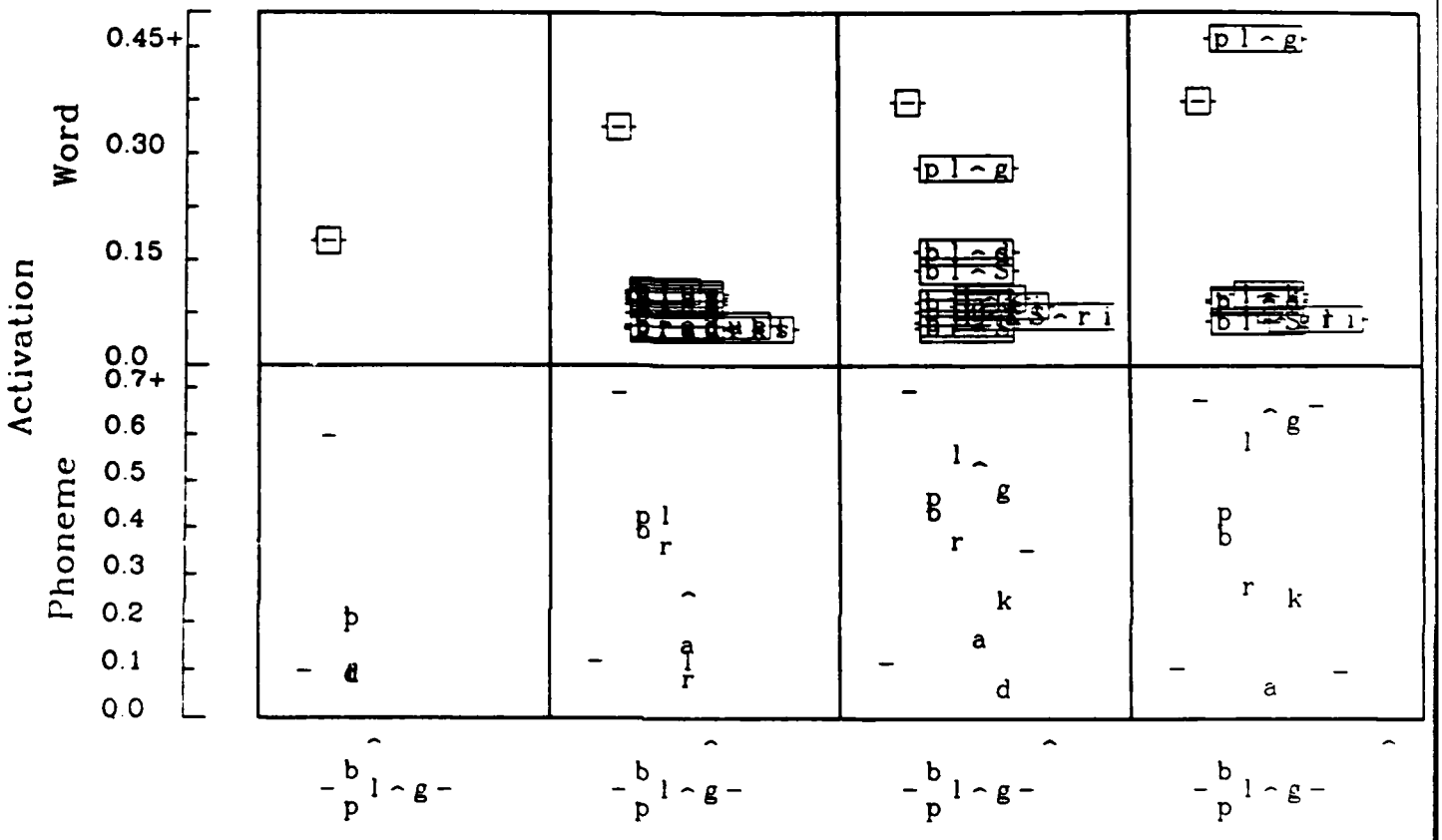


Figure 17. Phoneme and word level activations at several points in the unfolding of a segment ambiguous between /b/ and /p/, followed by /l/, /ˆ/, and /g/.

ous between /p/ and /b/; in one case, the ambiguous segment is followed by /l^g/ (as in "plug"; in the other, it is followed by /l^S/ ("as in blush". Thus, the lexical effect should lead to a /p/ response in the first case, but a /b/ response in the second case. The differences between the contexts do not begin to show up until after the center of the final phoneme, which occurs at slice 30. The reason for this is simply that the information is not available until that point.

Elimination of the lexical effect by time pressure. Recently, Fox (1982) has reported that the lexical effect on word initial segments is eliminated if subjects are given a deadline to respond within 500 msec of the ambiguous segment. Though they can correctly identify unambiguous segments in responses made before the deadline, these early responses show no sensitivity to the lexical status of the alternatives. From these findings, Fox argued that the lexical effect is a post-perceptual, "guessing" effect, rather than an actual effect of lexical information on perception.

Our model is completely consistent with Fox's results. Indeed, we have already seen that the activations in the TRACE only begin to reflect the lexical effect about one phoneme or so after the phoneme that establishes the lexical identity of the item. Given that this segment does not occur, in Fox's experiments, until the second or third segment after the ambiguous segment, there is no way that a lexical effect could be observed on early responses.

But what about the fact that early responses to unambiguous segments can be accurate? TRACE accounts for this too. In Figure 19 we show the state of the TRACE at various different points after the unambiguous /b/ in /bl^g/. Here, the /b/ dominates the /p/ from the earliest point. The analogous result is obtained, when the stimulus is /p/ in /pl^g/, and the activation for the initial phoneme is quite independent of whether or not the item is a word. The response strength for the case when /pl^g/ is presented in Figure 18 shows that the probability of choosing /p/ is near unity within 12 processing cycles, or 300 msec of the initial segment, well before the deadline would be

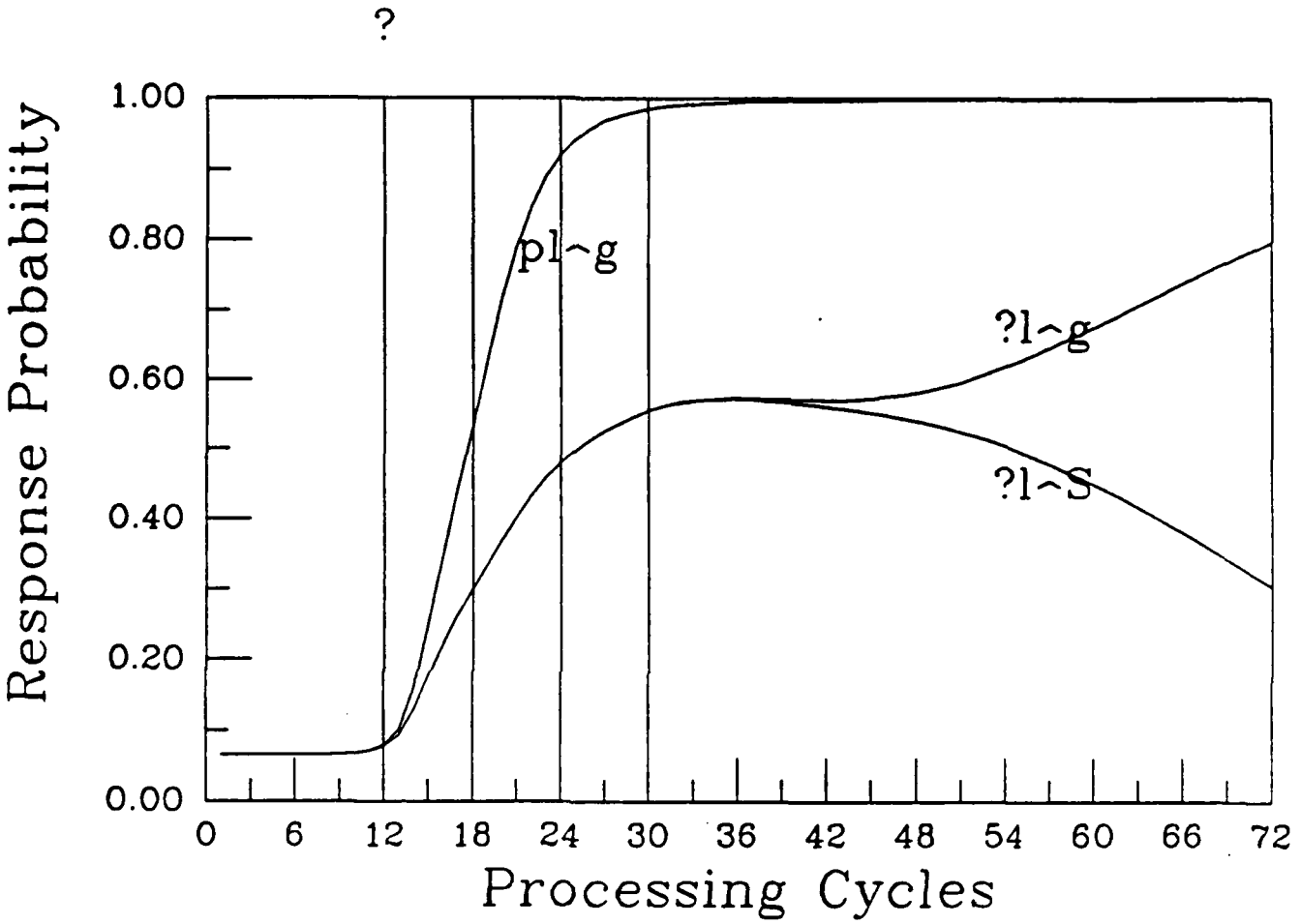


Figure 18. The time-course of the build up in the strength of the /p/ response based on activations of phoneme nodes in slice 12, in processing an ambiguous /b/ - /p/ segment in /_l'g/, and the same segment in /_l'S/. Also shown is the build-up of response strength for processing an unambiguous /p/ segment in /pl'g/. The vertical line topped with a *"?"* indicates the point in time corresponding to the center of the initial segment in the input stream. Successive vertical lines indicate centers of successive phonemes.

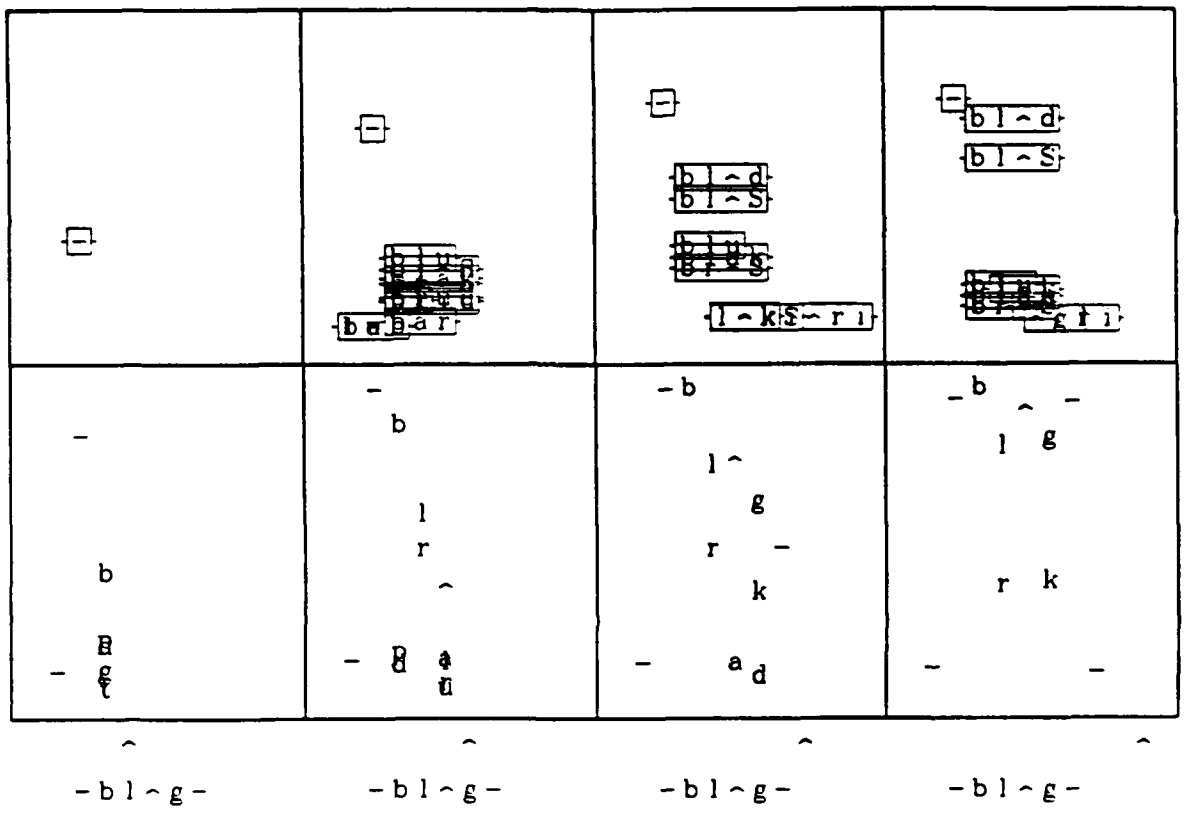


Figure 19. The state of the Trace at various stages of processing the stream /bl'g/.

reached, -- and well before word-identity specifying information is available.

Lexical effects late in a word. Lexical effects on word-initial segments develop rather late, at least in the case where there is no context preceeding the word. The lexical effect need not develop late in all cases, however. When the information that precedes the ambiguous segment has already established which of the two alternatives for the ambiguous segment is correct, TRACE shows a lexical effect that develops along with the direct perceptual information establishing the identity of the target segment. This phenomenon is illustrated in Figure 20, which shows the state of the Trace at several points in time relative to an ambiguous final segment that could be a /t/ or a /d/, at the end of the context /targ˚/. Within the duration of a single phoneme after the center of the ambiguous segment, /t/ already has an advantage over /d/. We therefore predict that Fox's results would come out differently, were he to use word-final, as opposed to word-initial, ambiguous segments. In such a case we would expect the lexical effect to show up, well within the 500 msec deadline.

Absence of lexical effect in monitoring word-initial phonemes. Foss and Blank (1980) presented some results which seemed to pose a challenge to interactive models of phoneme identification in speech perception. They gave subjects the task of listening to spoken sentences for occurrences of a particular phoneme in word-initial position. Reaction time to press a response key from the onset of the target phoneme was the dependent variable. In one example, the target was /g/ and the sentence was:

At the end of last year, the government ...

The subject's task was simply to press the response key upon hearing the /g/ at the beginning of the word government.

The principle finding of Foss and Blank's study was that it made no difference whether the target came at the beginning of a word or a nonword. From this, they argued that lexical influences

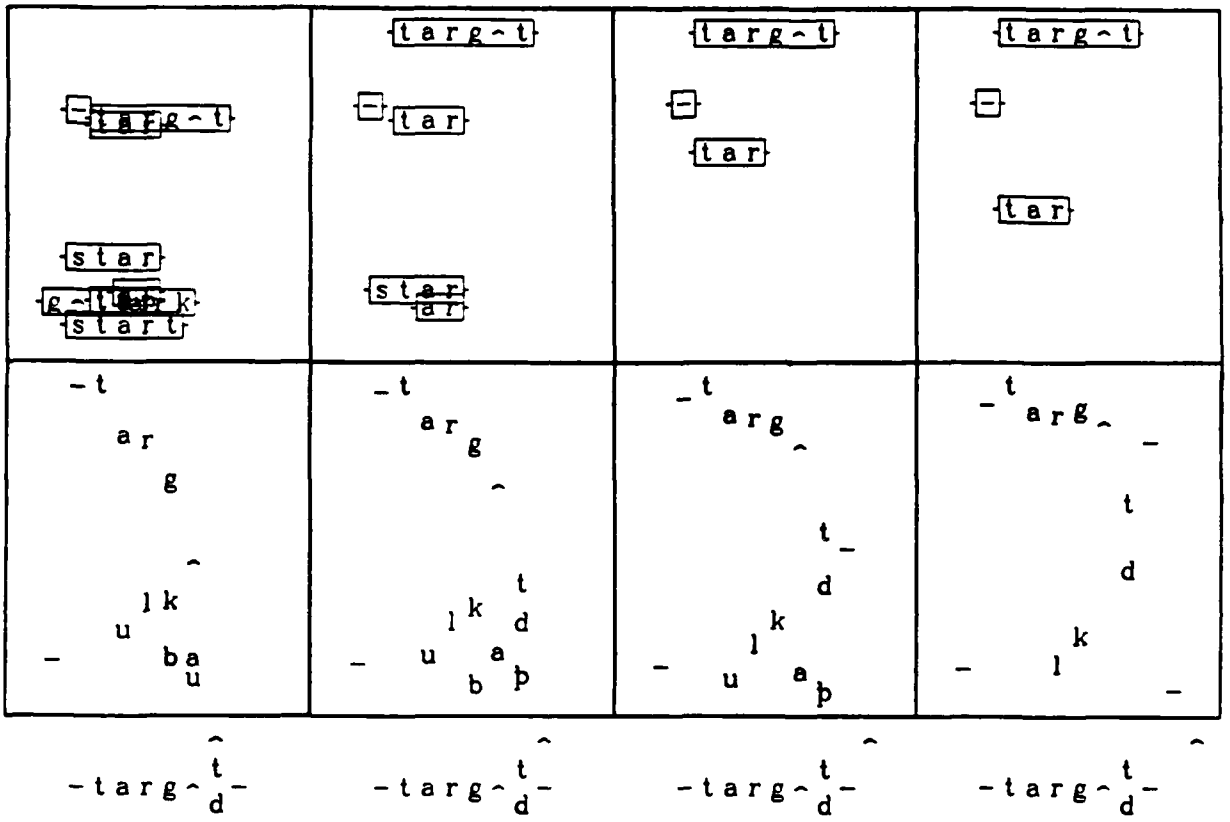


Figure 20. The state of the Trace at several stages of processing the stream consisting of /targ^/ followed by a segment ambiguous between /t/ and /d/.

show up in phoneme monitoring tasks only when the task becomes difficult; in the normal course of events, as long as phoneme identification is easy, lexical information is not used.

This conclusion seems partly at variance with the spirit of the TRACE model, since in TRACE, the lexical level is always involved in the perceptual process. However, we have already seen that there are conditions under which the lexical level does not get much of a chance to exert an effect. In the previous section we saw that there is no lexical effect on identification of ambiguous word-initial targets when the subject is under time pressure to respond quickly, simply because the subject must respond before information is even available that would allow the model -- or any other mechanism -- to produce a lexical effect.

In the Foss and Blank situation, there is even less reason to expect a lexical effect, since the target is not an ambiguous segment. We already saw that activation curves rise rapidly for unambiguous segments; in the present case, they can reach near-peak levels well before the acoustic information that indicates whether the target is in a word or nonword is "off the tape."

The results of a simulation run illustrating these points are shown in Figure 21. For this example, we imagine that the target is /t/. Note how during the initial syllable of both streams, little activation at the word level has been established. Even toward the end of the steam, where the information is just coming in which determines that "trugus" is not a word, there is little difference, because in both cases, there are several active word-level candidates, all supporting the word initial /t/. It is only after the end of the stream that a real chance for a difference has occurred. Well before this time arrives, the subject will have made a response, since the strength of the /t/ response reaches a level sufficient to guarantee a high accuracy by about cycle 30, well before the end of the word, as illustrated in Figure 22.

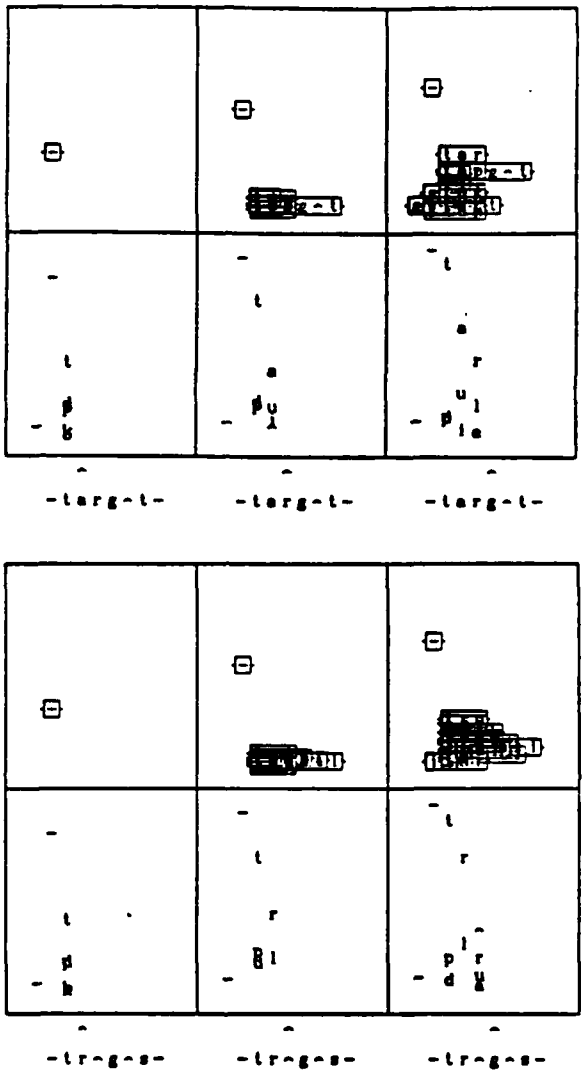


Figure 21. State of the Trace at three different points during the processing of the word "target" (/targ˙t/) and the nonword "trugus" (/tr˙g˙s/).

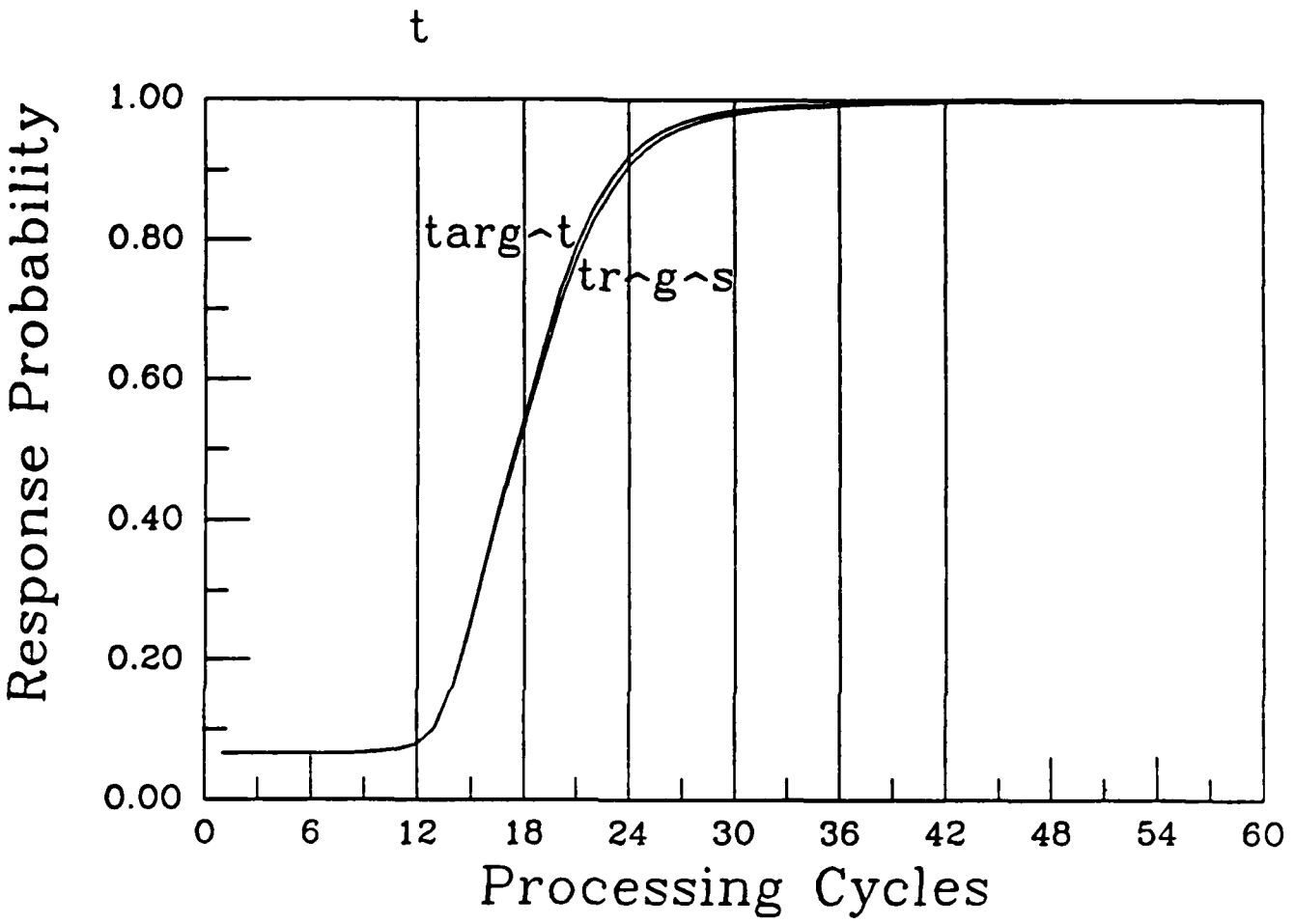


Figure 22. Time course of growth in the probability of the /t/ response based on activations of phoneme nodes in slice 12, during processing of /targ^t/ and /tr^g^s/. The vertical lines indicate the peaks on the feature patterns corresponding to the successive phonemes of the presented word.

Even though activations are quite rapid for unambiguous segments, these can still be influenced by lexical effects, provided that the lexical information is available in time. In Figure 23, we illustrate this point for the phoneme /t/ in the streams /sikr˚t/ (the word "secret") and /g˚ld˚t/ (a nonword). The Figure shows the strength of the /t/ response as a function of processing cycles, relative to all other responses based on activations of phoneme nodes at cycle 42, the peak of the input specification for the /t/. Clearly, response strength grows faster for the /t/ in /sikr˚t/ than for the /t/ in /g˚ld˚t/; picking an arbitrary threshold of .9 for response initiation, we find that the /t/ in /sikr˚t/ reaches criterion about 3 cycles or 75 msec sooner than the /t/ in /g˚ld˚t/.

Marslen-Wilson (1980) has reported an experiment that demonstrates the existence of lexical effects in phoneme monitoring for phonemes coming at different points in words. For phonemes coming at the beginning of a word or at the end of the first syllable, he found no facilitation for phonemes in words relative to phonemes in nonwords (in fact there was a nonword advantage for these early target conditions). For targets occurring at the end of the second syllable of a two-syllable word (like "secret" -- though the stimuli in this particular experiment were Dutch) Marslen-Wilson found a 85 msec advantage compared to corresponding positions in non-words. This compares quite closely with the value of about 75 msec we obtained for the /sikr˚t/-/g˚ld˚t/ example. At the ends of even longer words, the word advantage increased in size to 185 msec. Marslen-Wilson's result thus confirms that there are indeed lexical effects in phoneme monitoring -- even for unambiguous inputs -- but underscores the fact that there is no word advantage for phonemes whose processing can be completed long before lexical influences would have a chance to show up.

Are Phonotactic Rule Effects the Result of a Conspiracy?

Recently, Massaro and Cohen (1983) have reported evidence supporting the use of phonotactic rules in perception. In one experiment, Massaro and Cohen's stimuli consisted of phonological seg-

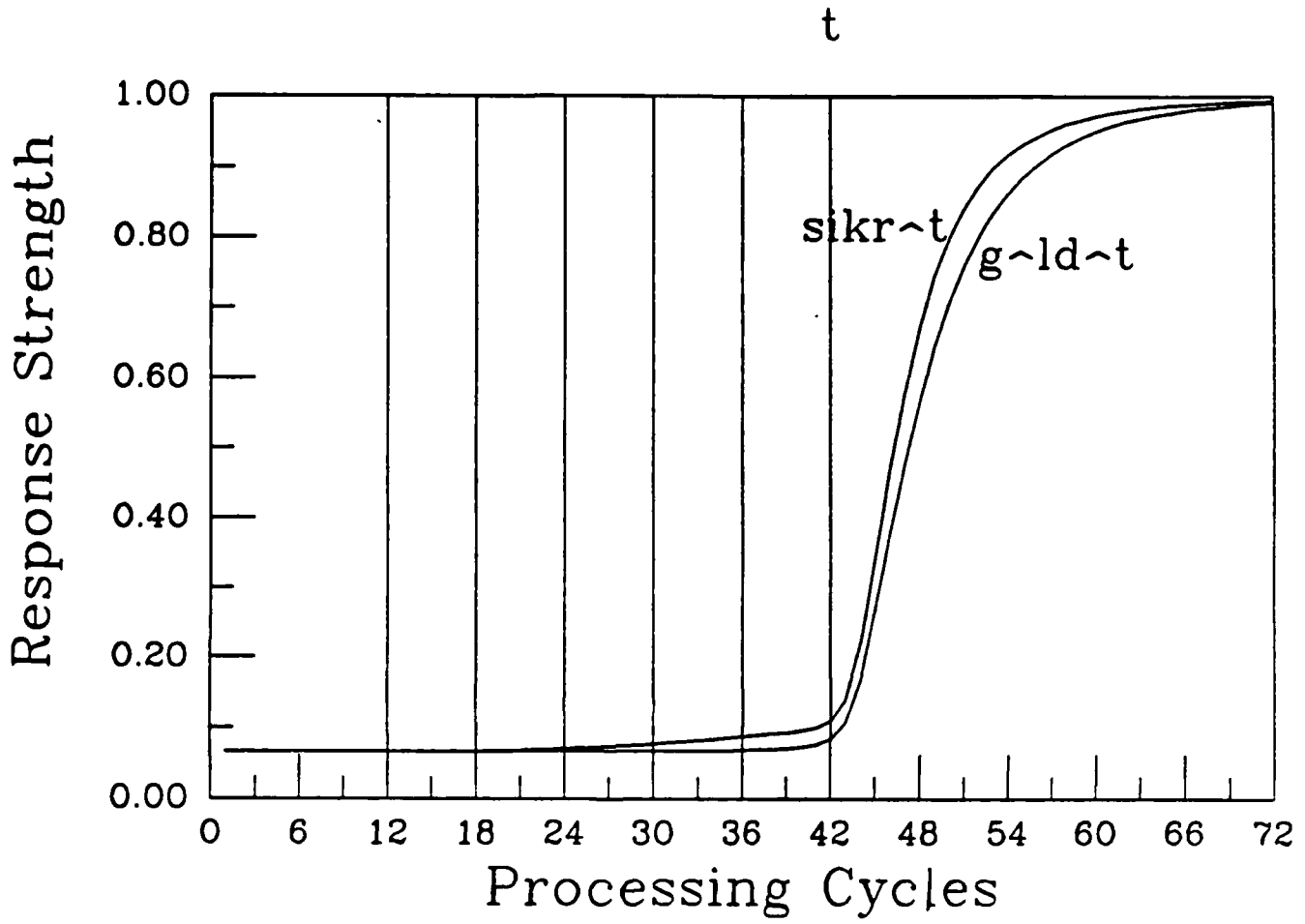


Figure 23. Probability of the /t/ response as a function of processing cycles, based on activation of phoneme nodes at cycle 42, for the stream /sikr˜t/ ("secret") and /g˜ld˜t/ ("guldut"). Vertical lines indicate the peaks of the input patterns corresponding to the successive phonemes in either stream.

ments ambiguous between /r/ and /l/ in different contexts. In one context (/t_i/) /r/ is permissible in English, but /l/ is not. In another context (/s_i/) /l/ is permissible in English but /r/ is not. In a third context (/f_i/) both are permissible, and in a fourth (/v_i/) neither is permissible. Massaro and Cohen found a bias to perceive ambiguous segments as /r/ when /r/ was permissible; or as /l/ when /l/ was permissible. No bias appeared in either of the other two conditions.

With most of these stimuli, phonotactic acceptability is confounded with the actual lexical status of the item; thus /fli/ and /fri/ (flee and free) are both words, as is /tri/ but not /tli/. In the s_i context, however, neither /sli/ or /sri/ are words, yet Massaro and Cohen found a bias to hear the ambiguous segment as /l/, in accordance with phonotactic rules.

It turns out that the TRACE model produces the same effect, even though it lacks phonotactic rules. The reason is that the ambiguous stimulus produces partial activations of a number of words (sleep and sleet in the model's lexicon; as well as sleeve, sleek, and others in the full lexicon of English words). None of these word units gets as active as it would if the entire word had been presented. However, all of them (in the simulation, there are only two, but the principle still applies) are partially activated, and all conspire together and contribute to the activation of /l/. This feedback support for the /l/ allows it to gain the upper hand over the /r/, just as it would if /sli/ were an actual word, as shown in Figure 24.

The hypothesis that phonotactic rule effects are really based on word activations leads to a prediction: That we should be able to reverse these effects if we present items that are supported strongly by one or more lexical items even if they violate phonotactic rules. A recent experiment by Elman (in preparation) confirms this prediction. In this experiment, ambiguous phonemes (for example, halfway between /b/ and /d/) were presented in three different types of contexts. In all three types, one of the two (in this case, the /d/) what phonotactically acceptable, while the other (the /b/) was not. How-

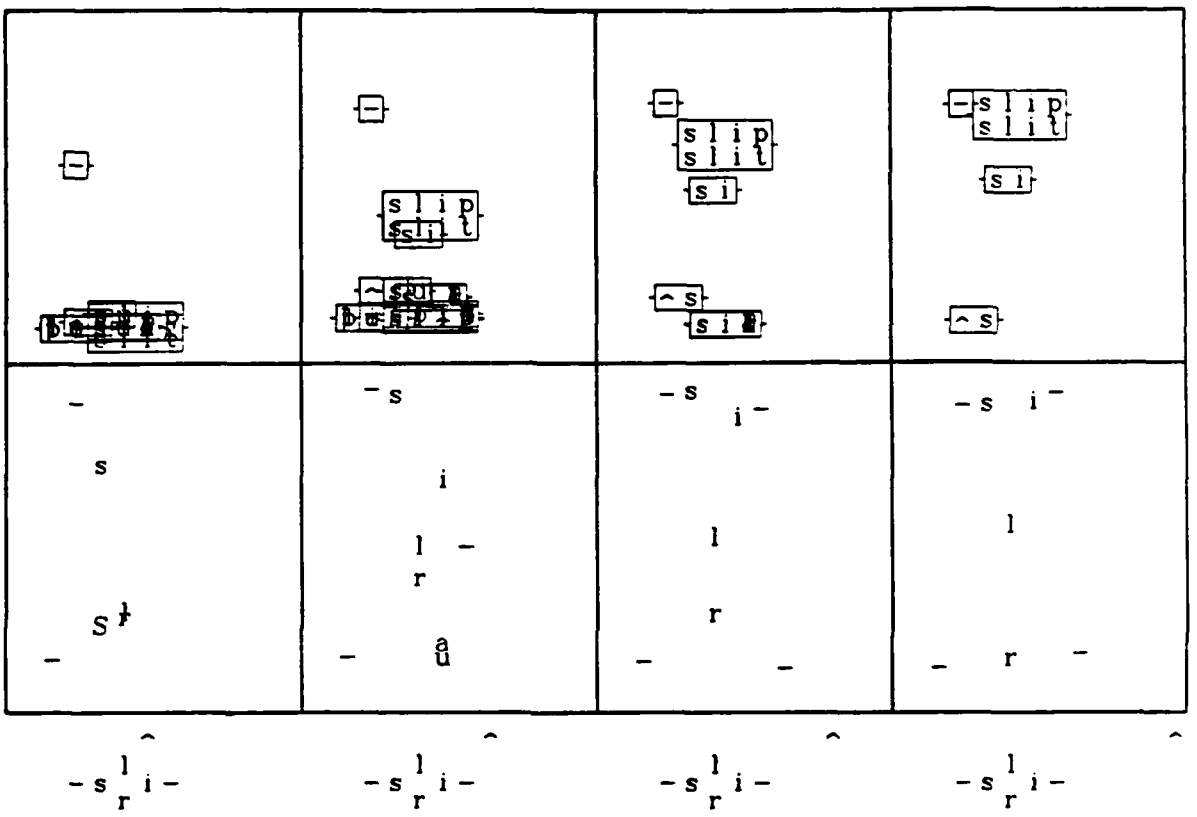


Figure 24. State of the Trace at several points in processing a segment ambiguous between /l/ and /r/, in the context /s_i/. The nodes for /slip/ and /slit/ are boxed together since they take on identical activation values.

ever, the contexts differed in their relation to words. In one case, the legal item actually occurred in a word ("bwindle"-"dwindle"). In a second case, neither item made a word, but the illegal item was very close to a word ("bwacelet"-"dwacelet"). In a third case, neither item was particularly close to a word ("bwiffle"-"dwiffle"). Results of the experiment are shown in Table 3. The presence of a word similar or identical to one of the two alternatives strongly influenced the subjects' choices between the two alternatives. Indeed, in the case where the phonotactically irregular alternative was very similar to a particular lexical item ("blacelet"), subjects tended to hear the ambiguous item in accord with the lexicon, even though it was phonotactically incorrect.

To determine whether the model would also produce such a reversal of the phonotactic rule effects with the appropriate kinds of stimuli, we ran a simulation using a simulated input ambiguous between /p/ and /t/ in the context /_luli/. /p/ is phonotactically acceptable in this context, but /t/ in this context makes an item that is very close to the word "truly". The results of this run, at two different points during processing, are shown in Figure 25. Early on in processing, there is a slight bias in favor of the /p/ over the /t/, because at first a large number of /pl/ words are slightly more activated than any words beginning with /t/. Later, though, the /t/ gets the upper hand, as the word "truly" comes to dominate at the word level. Thus, by then end of the word or shortly thereafter, the closest word has begun to play a dominating role, causing the model to prefer the phonotactically inappropriate interpretation of the ambiguous initial segment.

Of course, at the same time the word truly tends to support /r/ rather than /l/ for the second segment. Thus, even though this segment is not ambiguous, and the /l/ would suppress the /r/ interpretation in a more neutral context, the /r/ stays quite active.

Table 3

Percent Choice of Phonotactically Irregular Consonant

Stimulus type	Example	Percentage of identifications as "illegal" phoneme
legal word/illegal non-word	dwindle/dwindle	37%
legal non-word/illegal non-word	dwiffle/bwiffle	46%
legal non-word/illegal near-word	dwacelet/bwacelet	55%

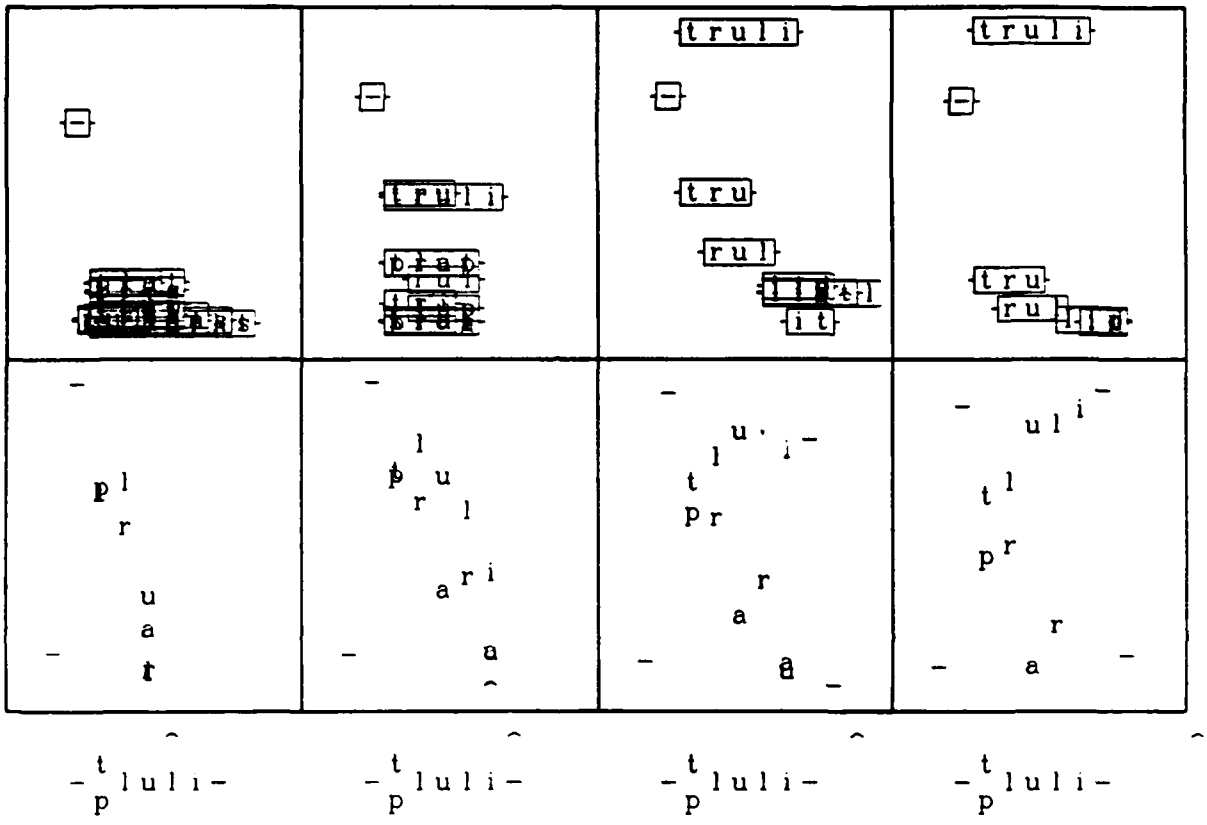


Figure 25. State of the Trace at several points in processing an ambiguous /p/-/t/ segment followed by /luli/.

Recognition of Words in Isolation

In a sense, we have gotten a bit ahead of ourselves in the last two sections, since we have considered how feedback from the word level influences phoneme identification, without considering the process of word activation itself. This section and the next are devoted to this topic. In this section we consider the recognition of words in isolation (i.e., surrounded by silence). In the next section we consider the recognition of words embedded in longer sound streams and the parsing of multiword utterances into separate words.

Computational issues. In considering the problem of word recognition, a paradox arises. On the one hand, our perceptual system appears to rely heavily on the beginning parts of words. People are more likely to stumble over slight mispronunciations in beginning parts of words, rather than fluently restoring them, and when listening to running speech explicitly for mispronunciations, those that occur in early parts of words are more easily missed. On the basis of this and other evidence, Marslen-Wilson and Welsh (1978) proposed a model in which word recognition involves three basic assumptions. 1) The model uses the first sound of the word to determine which words will be in the initial cohort or candidate set. 2) Once the candidate set is established, the model eliminates words as each successive phoneme arrives if the new phoneme in the input fails to match the next phoneme in the word. 3) Word recognition occurs when the cohort has been reduced to a single member; in an auditory lexical decision task, the decision that an item is a non-word can be made as soon as there are no remaining members in the cohort. In other papers Marslen-Wilson has provided impressive evidence consistent with the cohort model (Marslen-Wilson, 1980). In general, this evidence supports what we call the immediacy principle, which states that word identification processes are nearly immediately sensitive to each new aspect of the speech stream as it arrives at the ear of the perceiver.

Limitations of the Cohort Model. Though the Cohort model captures the immediacy principle quite nicely and represents an important source of the inspiration of the present modeling effort, there are computational difficulties with it. As Thompson (1984) has pointed out, the Cohort model relies heavily on word beginnings, and there is no explicit procedure described for recovering words into the cohort once they have been excluded from it. This makes it very difficult to understand why it is so easy to recognize words with distorted (e.g., "dwibble") or noise-obliterated (e.g., #lippery) beginnings. These kinds of examples might be taken as suggesting that what we need to do is liberalize the criterion for rejecting words from inclusion in the cohort. This would allow mild distortions like replacing /r/ with /w/ not to disqualify a word from the cohort.

The problem with this scheme is that sometimes we want to be able to eliminate items which mismatch the input on a single feature and sometimes we do not. Consider the items "pleasant" and "blacelet". In the first case, we need to exclude "present" from the cohort, so the slight difference between /l/ and /r/ must be sufficient to rule it out; in the second case, we do not want to loose the word "bracelet", since it provides the best fit overall to the input. Thus, in this case, the difference between /l/ and /r/ must not be allowed to rule a word candidate out.

Thus the dilemma: On the one hand, we want a mechanism that will be able to recognize words shortly after the input makes them unique, to account for Marslen-Wilson's results, and to capture the computationally important immediacy principle. On the other hand, we do not want the model to eliminate possibilities which might later turn out to be correct.

Thompson (1984) has also pointed out another problem with the Cohort approach. Its reliance on word beginnings makes it unsuited for dealing with cases in which there is ambiguity about where one word ends and the next begins. An adequate computational model of speech perception, then, must not only capture both horns of the dilemma raised above; it must also deal with the indeter-

minacy of word beginnings.

In this section and the next we will show that TRACE goes a long way toward meeting both of these criteria, by virtue of the simple workings of the activation and competition mechanisms. The present section considers isolated word recognition, where the word boundary problem does not arise; the next section considers the case in which the model must decide, not only what the words are, but where they begin and end.

Competition vs bottom-up inhibition. TRACE deals with the dilemma posed above by using competition, rather than phoneme-to-word inhibition. The essence of the idea is simply this. Phoneme nodes have excitatory connections to all the word nodes they are consistent with. Thus, whenever a phoneme becomes active in a particular slice of the Trace, it sends excitation to all the word nodes consistent with that phoneme in that slice. The word nodes then compete with each other to see which one provides a better overall fit to the pattern of activation at the phoneme level. Consider, from this point of view, our two items "pleasant" and "blacelet" again. In the first instance, "pleasant" will receive more bottom-up excitation that "present", and so will win out in the competition. We have already seen, in our analysis of categorical perception at the phoneme level, how even slight differences in initial bottom-up excitation can be magnified by the joint effects of competition and feedback. But the real beauty of the competition mechanism is that its action is contingent on the activation of other word candidates. Thus, in the case of "blacelet", since there is no word "blacelet", "bracelet" will not be suppressed. Initially, it is true, words like "blame" and "blatant" will tend to dominate "bracelet", but since the input matches "bracelet" better than any other word, it will eventually come to dominate the other possibilities.

This behavior of the model is illustrated using examples from its restricted lexicon in Figure 26. In one case, the input is "legal", and the word "regal" is dominated. In the other case, the input is

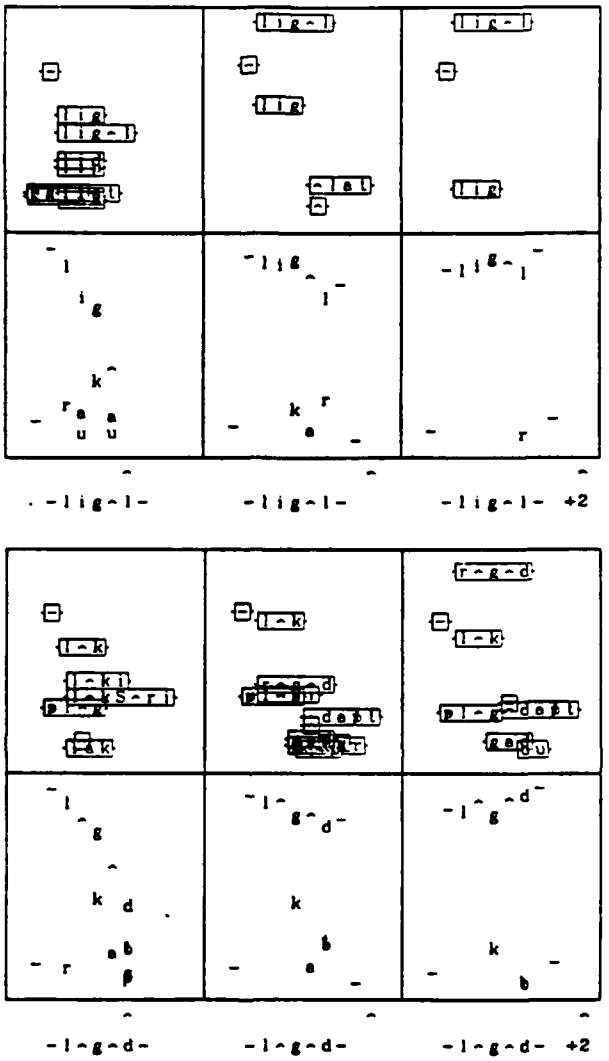


Figure 26. State of the Trace at two points during processing of "legal" and "lugged".

"lugged", and the word "rugged" tends to dominate. "rugged" must compete with other partial matches of "lugged", of course, and it is less effective in this regard than it would be if the input exactly matched it.

It should be noted that the details of what word will win in such cases depend on a number of factors. For example, as we shall see, it is possible to find cases in which a word the correctly spans a part of a longer string dominates a longer word that spans the whole string but misses out on a phoneme in one place or another. An item like "vigorette" may or may not be a case in point. In such cases, though, the most important thing might not turn out to be winning and losing, but rather the fact that both tend to stay in the game to a certain extent. Such neologisms can suggest a poetic conjunction of meanings, if used just right: "He walked briskly down the street, puffing his vigorette."

Time course of word recognition in TRACE. So far we have shown how TRACE overcomes a difficulty with the Cohort model, but we have not shown how well it can do at capturing the positive features of Cohort, namely that a winner begins to emerge very shortly after the information which specifies it comes in. We say "begins to emerge" in TRACE, because there is nothing absolute or hard and fast about one word winning out over another. One word's activation may dominate another, but the TRACE itself makes no final decisions. However, a decision mechanism, looking at the TRACE, might be able to select one word in preference to others as soon as the relative response strength of one of the words exceeded some criterion. Such a decision mechanism might entertain a candidate set, such as "all the words beginning at time-slice 3", and might "select" one of these as its response as soon as its probability of being correct reaches some criterion. It is important, though, to realize that such a decision would not necessarily in and of itself alter the Trace in any way. At a later point in time, with more information, the state of the TRACE might have changed based on new input. In such a case, the model would allow for the possibility of making different decisions at different times.

But our theorizing has somewhat outstripped our data at this point. Indeed, there is little very direct data bearing on the issue of when a subject can decide a particular word has been presented. In the absence of data, it is interesting to see how close TRACE comes to emulating the cohort model under conditions of undistorted input. To examine this, we considered the processing of the word "product" (/praduct/). Figure 27 shows the state of the Trace a various points in processing the word "product", and Figure 28 shows the response strengths of several nodes relative to the strength of the word "product" itself, as a function of time relative to the arrival of the successive phonemes in the input. In this Figure, the response strength of product is simply set to 1.0 at each time slice and the activations of nodes for other words are plotted relative to the activation of "product". The curves shown are for the word "trot", "possible", "priest", "progress", and "produce"; these words differ from the "product" (according to our encoding of them!) in the first, second, third, fourth and fifth phonemes respectively. On the basis of the immediacy principle, we expect the response strengths of a word to begin to drop off very shortly after the input begins to diverge from the word. Figure 28 illustrates that this is exactly what happens; each curve begins to drop off within about two slices (50 msec) of the peak of the first divergent phoneme. This is just the kind of behavior the Cohort model would produce in this case.

There is one aspect of TRACE's behavior which differs from Cohort: Among those words that are consistent with the input up to a particular point in time, TRACE shows a bias in favor of shorter words over longer words. Thus, "priest" has a slight advantage before the /a/ comes in, and "produce" is well ahead of "product" until the /^/ comes in. We do not know whether human subjects show this same sort of bias; it might be possible, if somewhat tricky, to test this aspect of the model using cross-modal priming experiments.

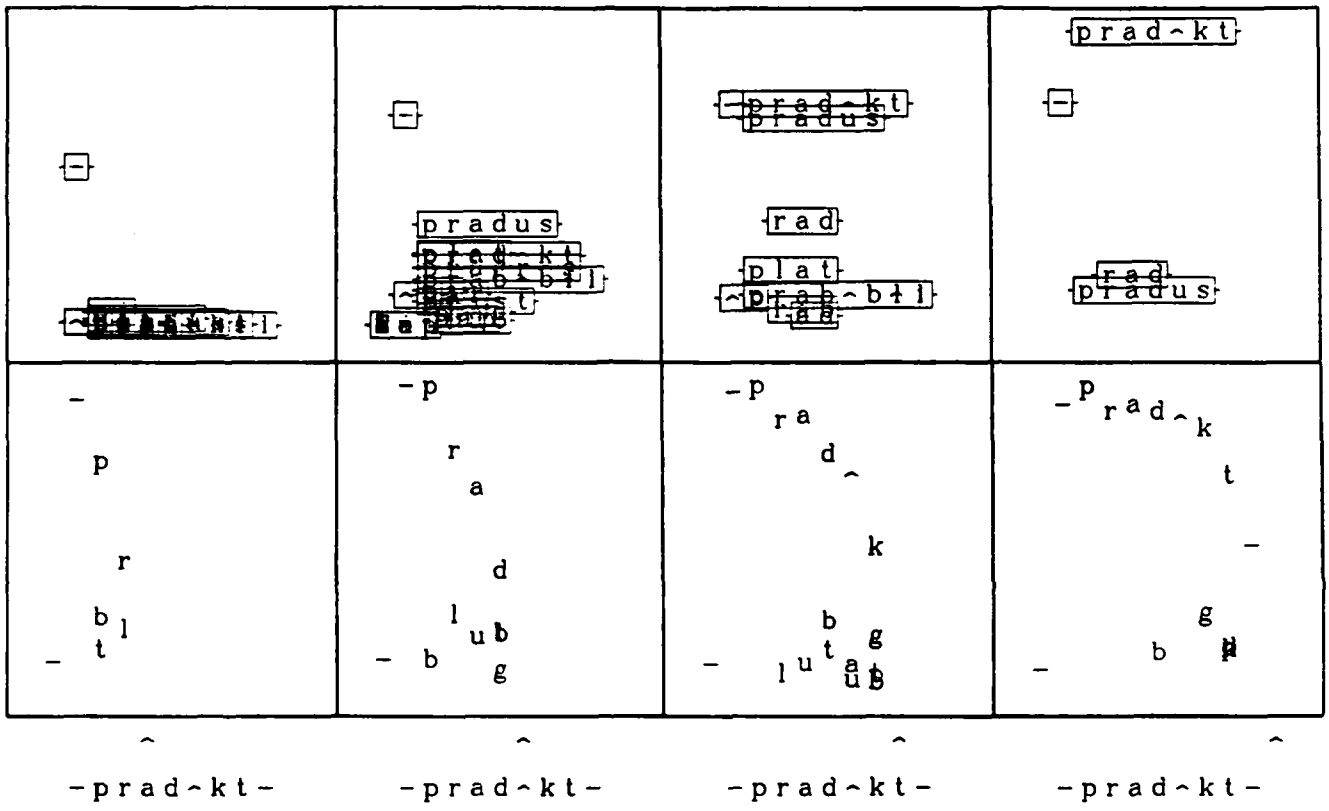


Figure 27. State of the Trace at various points in processing the word "product".

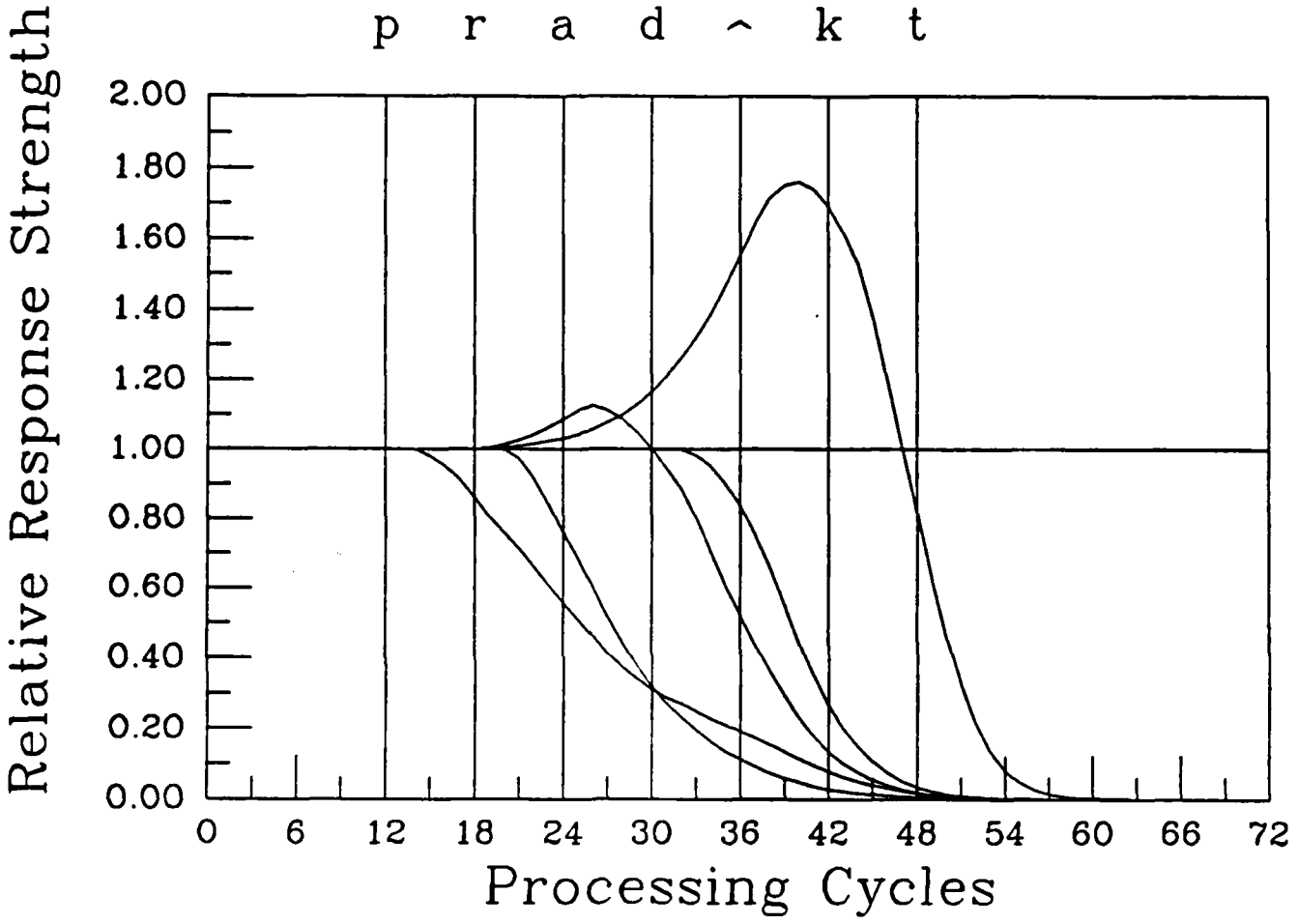


Figure 28. Activations of the nodes for several words relative to the activation of the node for "product", as a function of time relative to the peak of the first phoneme that fails to match the word. The successive curves coming off of the horizontal line representing the normalized activation of "product" are for the words "trot", "possible", "priest", "progress", and "produce", respectively. In our lexicon they are phonetized /trat/, /pas^b^l/, /prist/, /pragr^s/, and /pradus/ respectively.

TRACE's success in improving on the Cohort model owe to the use of competition instead of bottom-up inhibition. Competition allows a winner to dominate other alternatives, thereby capturing the benefits of bottom-up inhibition; while at the same time avoiding the pitfall of excluding near-misses from consideration when there is no better match.

Lexical Basis of Word Segmentation

How do we know when one word ends and the next word begins? This is by no means an easy task. Speech is unlike writing, in that there are no gaps between words within the same breath group. There are some cues. Certain phonemes are pronounced differently in word initial than in medial positions; certain kinds of coarticulatory effects are blocked at word boundaries (nitrate is pronounced slightly differently than night rate); word initial phonemes are often lengthened; and there are sometimes other cues as well. Doubtlessly, we exploit these cues in speech perception. But quite frequently, particularly in fluent speech, there is nothing to tell us where one word ends and the next one begins other than our knowledge of words and the sense we can make out of sequences of them. Our present model lacks syntactic and semantic levels, so it cannot make use of these higher-level constraints; but it can make use of its knowledge about words, not only to identify individual words in isolation, but to pick the sequence of words out of a stream of running speech. Word segmentation emerges from the interactive activation process, as part and parcel of the process of word perception.

This section considers several aspects of the way in which word segmentation emerges form the interactive activation process, as observed in simulations with TRACE II. Before we consider these, it is worth recalling the details of some of the assumptions made about the bottom up activation of word nodes and about competitive inhibition between word nodes. First, the extent to which a particular phoneme activates a particular word node is independent of the length of the word. This means that, when all the letters of a longer word are present, the word will receive more bottom-up support than

a shorter word that is a part of the longer word, simply because the longer word is receiving support from a larger number of phonemes. Second, the extent to which a particular word node will inhibit another word node is proportional to the temporal overlap of the two word nodes. This means that words which span non-overlapping stretches will not inhibit each other, but will gang up on other words than partially overlap each of them. These two assumptions form most of the basis of the effects we will observe in the simulations.

The boundary is in the ear of the behearer. First, we consider the basic fact that the number of words we will hear in a sequence of phonemes can depend on our knowledge of the number of words the sequence makes. Consider the two inputs /barti/ and /parti/. Though we can say either item in a way that makes it sound like a single word or like two words, there is an intermediate way of saying them so that the first sounds like two words (perhaps the brand of a cattle ranch?) and the other like only one (to get this effect it helps to pretend one is a computer speaking).

To see what TRACE II would do with these and other cases, we ran simulation experiments with each individual word in the lexicon preceeded and followed by silence, and then with 211 pairs of words, with a silence at the beginning and at the end of the entire stream. The pairs were made by simply permuting the lexicon twice and then abutting the two permutations so that each word occurred once as the first word and once as the second word in the entire set of 211 pairs.

With the individual words, TRACE made no mistakes -- that is, by a few slices after then end of the word, the word that spanned the entire input was more strongly activated than any other word. An example of this is shown using the item /parti/ in Figure 29. The steam /parti/ might be either one word or two ("par tea", or "par tee"). As it happens, however, TRACE tends to treat it as one, unless a pause is interposed between the /par/ and the /ti/, as illustrated in Figure 30.

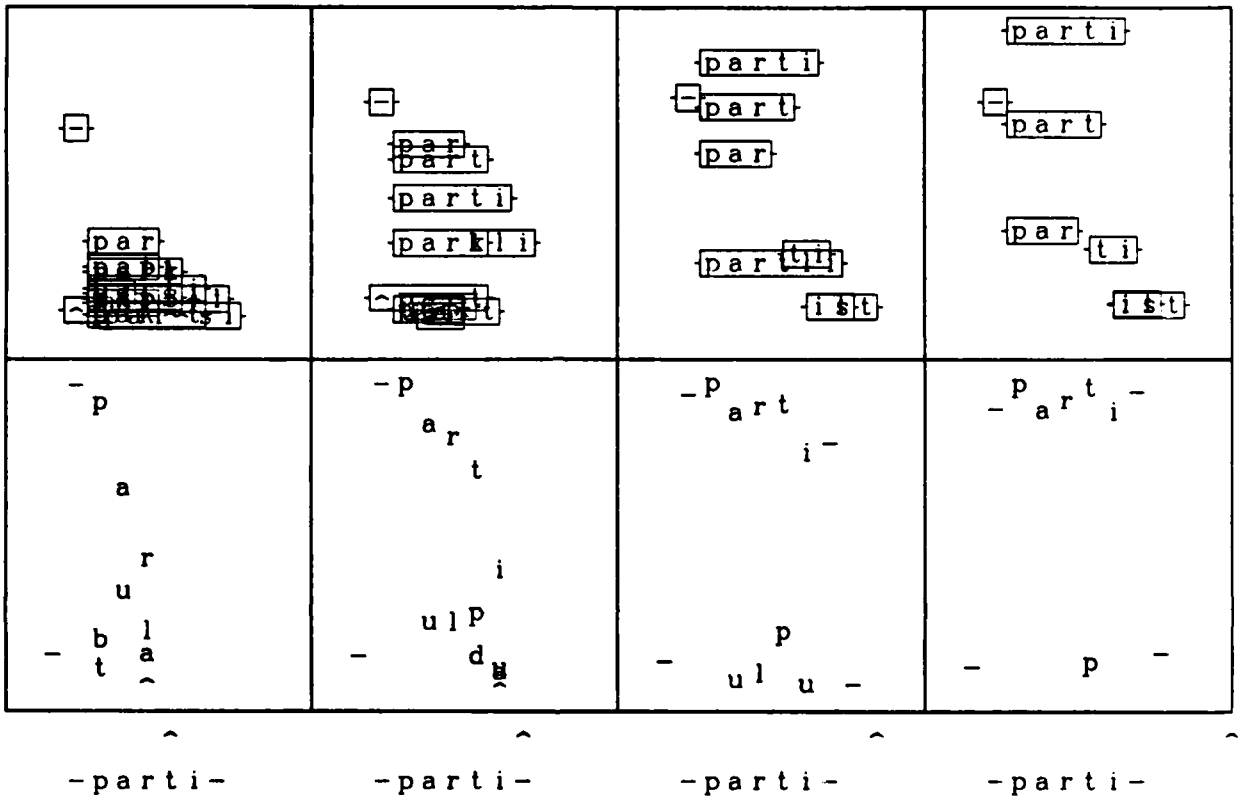


Figure 29. The state of the Trace at various points during processing of /parti/.

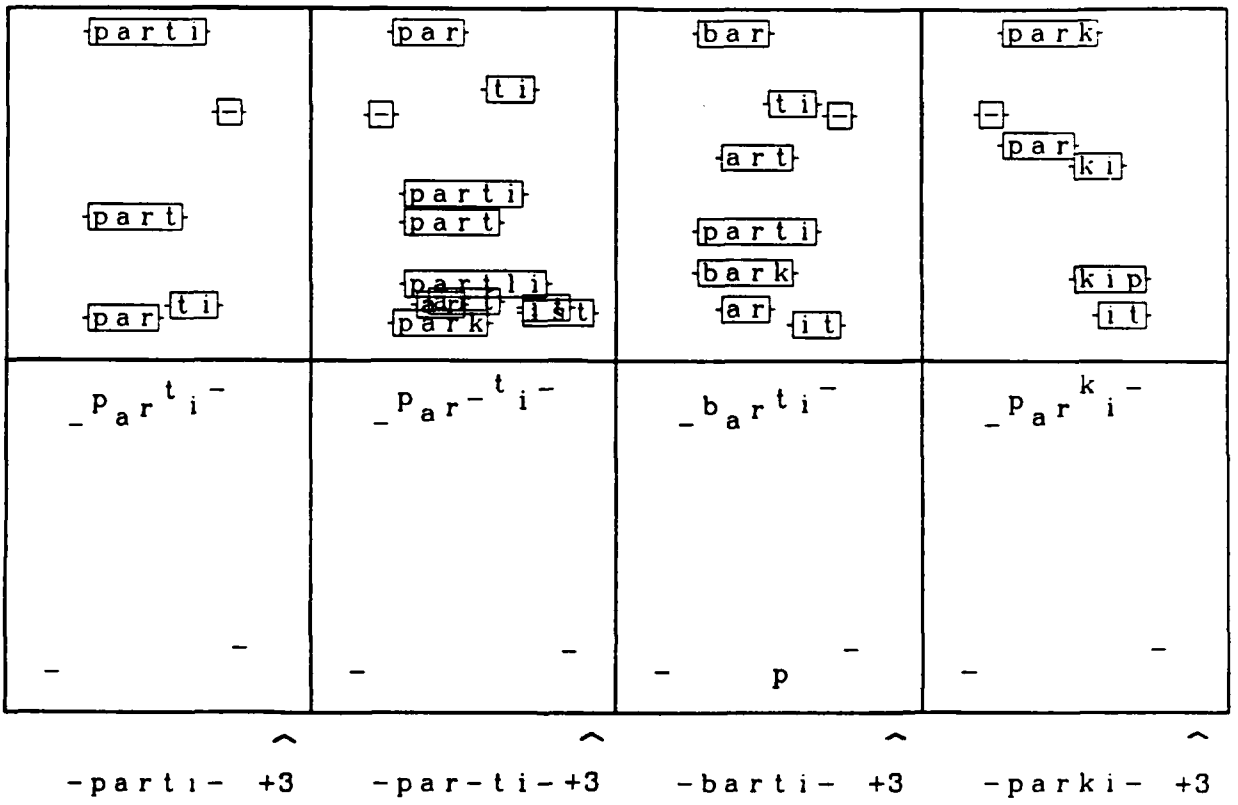


Figure 30. State of the Trace after processing the streams /parti/, /par-ti/, /barti/, and /parki/.

Why do longer words tend to win out over two shorter ones in TRACE? There are two main reasons. First of all, a longer word receives more bottom-up support than either shorter word, as we have already noted, simply because there are more phonemes activating the longer word than the shorter word. The second reason has to do with the sequential nature of the input. In the case of /parti/, by the time the /ti/ is coming in, the word "party" is so well established that it keeps /ti/ from getting as strongly activated as it would otherwise, as illustrated in Figure 29. This behavior of the model leads to the prediction that short words imbedded in the ends of longer words should not get as strongly activated as shorter words coming earlier in the longer word. This prediction can probably be tested using a cross-modal priming paradigm such as the one used by Swinney (1982).

What happens when there is no long word that spans the entire stream, as in /barti/? In this case, the model settles on the two word interpretation "bar tee", As shown in Figure 30. Note that other words, such as "art", that span only a portion of the input, are less successful than either "bar" or "tee". The reason is that "bar" suffers two letters' competition from "art" but none from "tee", and "tee" suffers one letter's competition from "art" and none from "bar"; while "art" suffers a total of three letters competition, two from "bar" and one from "tee".

These remarkably simple mechanisms of activation and competition do a very good job of word segmentation, without the aid of any syllabification, stress, phonetic word boundary cues, or semantic and syntactic constraints. In 187 of the 211 word pairs tested in the simulation experiment, the model came up with the correct parse, in the sense that no other word was more active than either of the two words that had been presented. Some of the failures of the model occurred in cases where the input was actually consistent with two parses, either a longer spanning word rather than a single word (as in party) or a different parse into two words, as in "part rust" or "par trust". In such cases TRACE tends to prefer parses in which the longer word comes first. There were, however, some cases in which the model did not come up with a valid parse, that is a pattern than represents complete

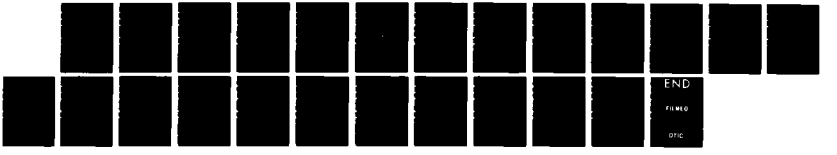
coverage of the input by a set of non-overlapping words. For example, consider the input /parki/. Though this makes the two words "par" and "key", the word "park" has a stronger activation than either "par" or "key", as illustrated in Figure 30.

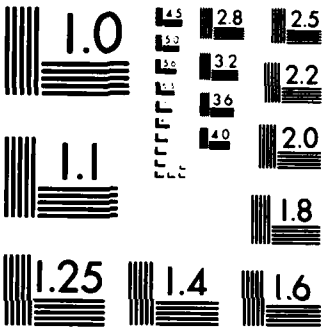
Whether this behavior is characteristic of the human perceiver is not clear. What is clear is that a complete model would also exploit syllabification, stress, and other cues to word identity to help eliminate some of the possible interpretations of TRACE II's simple phoneme streams. The activation and competition mechanisms in TRACE II are sufficient to do quite a bit of the word segmentation work, but we do not expect them to do this perfectly without the aid of other cues.

Actually, the fact that TRACE does not insist on an interpretation in which each phoneme is covered by one and only one word is often a virtue, since in many cases, the last phoneme of a word must do double duty as the first phoneme of the next, as in "hound dog" or "brush shop". While we tend to pronounce both identical consonants or to signal the doubling by lengthening a single consonant in careful speech, in rapid speech the differences between double and single consonants across word boundaries often disappear (as in /bustap/ -- is it "bus top" or "bus stop?"). The model is actually quite happy with overlapping words, even when a non-overlapping parse is available as with /bˆstap/, as shown in Figure 31.

One last example of TRACE II's performance in segmenting words is illustrated in Figure 32. The Figure shows the state of the Trace at several points during the processing of the stream /SiSˆtˆbaks/. By the end, the words of the phrase "She shut a box", which fits the input perfectly with no overlap, dominate all others. This simulation illustrates the special difficulty faced by short words like "a" in the model. Such words are almost always dominated by a longer spanning word, until, as in this case, subsequent input rules out the spanning word and allows the short word to emerge. (If the model exploited frequency weighting or syntactic and semantic constraints, it is likely that the bulk of

AD-A148 593 INTERACTIVE ACTIVATION MODEL OF SPEECH PERCEPTION(U) 2/2
CALIFORNIA UNIV SAN DIEGO LA JOLLA
J L MCCLELLAND ET AL. 01 NOV 84 N00014-82-C-0374
UNCLASSIFIED F/G 17/2 NL





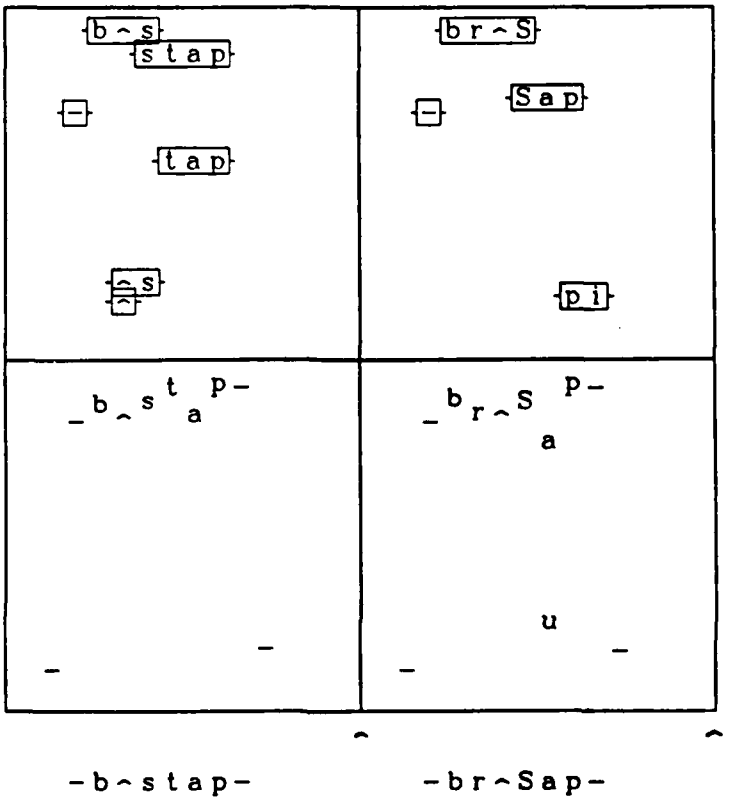


Figure 31. State of the Trace at the end of the streams /bustap/ ("bus stop" or "bus top") and /bruSop/ ("brush shop").

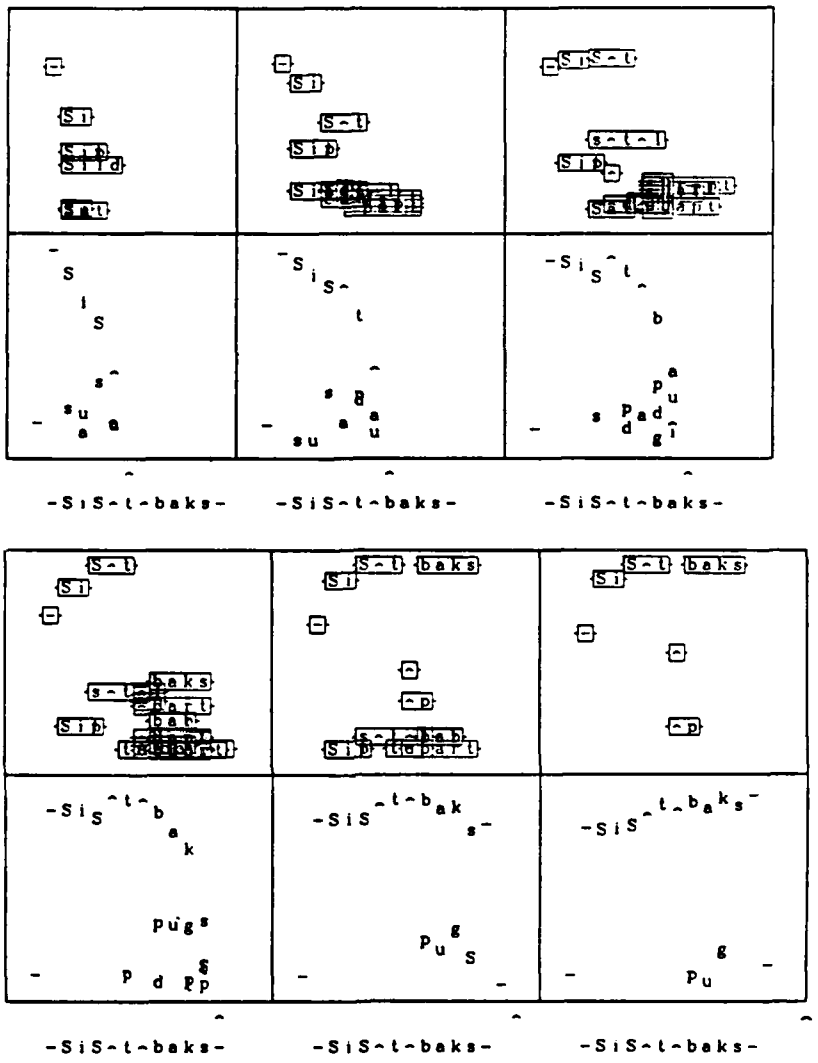


Figure 32. The state of the Trace at several points during the processing of the stream /SiS^t^baks/ ("She shut a box").

this problem would be eliminated, at least with words like "a" and "the".)

These examples show that the word activation/competition mechanism can go a long way toward providing a complete interpretation of the input stream as a sequence of words. It tends to prefer to interpret streams of phonemes as single longer words rather than as a sequence of short words; and it tends to find parses that account for each phoneme once; but it does not insist upon this, and will occasionally produce an interpretation that leaves part of the stream of phonemes unaccounted for or which accounts for part of the stream of phonemes twice. Often enough, it will also leave certain alternatives to its "preferred parse" in a strong position, so that both alternatives are made available to higher levels.

There is little evidence bearing directly on the details of these aspects of TRACE II's behavior. However, we will consider two empirical findings which are consistent with the model.

Where does a nonword end? If we are listening to speech, and someone suddenly interjects a nonword into it, it may be relatively easy to tell where the non-word begins, since it will be right after the end of the preceeding word. In most cases, the interactive activation process might well have established where the preceeding word ends at around the time of the end of the preceeding word, particularly if activations at the word level are aided by syntactic and semantic constraints. However, it is much harder to establish the end of a nonword, since the fact that it is a nonword means that we cannot exploit any knowledge of where it should end to do so.

This fact may account for the finding of Foss and Blank (1980) that subjects are much slower to respond to word-initial target phonemes at the beginning of a word preceeded by a nonword than at the beginning of a word preceeded by a word. For example, responses to detect word initial /d/ were faster in stimuli like the following

At the end of last year, the government decided ...

than they were when the word preceeding the target (in this case government) was replaced by a non-word such as "gatabont". The fact that reaction times were slower in this case suggests that subjects were able to use their knowledge of where one word ends to help them determine where the next word begins.

An example of how this would come about in TRACE is illustrated in Figure 33. In the example, the model receives the stream "possible target" or "pagusle target", and we imagine that the target is word-initial /t/. In the first case, the word "possible" is clearly established and competitors underneath it have been completely crushed by the time the /t/ becomes active at the phoneme level, so there is no ambiguity about the fact that this /t/ is at the beginning of the next word. In the second case, words beginning and ending at a number of different places are partly activated. Thus, the subject must wait until he is well into the word "target" before it becomes clear that the first /t/ in target is in fact a word-initial /t/.

In reality, the situation is probably not as bleak for the perceiver as it appears in this example, because in most cases there will be clues in the manner of pronunciation and the syllabification of the streams that will help to indicate the location of the word boundary. However, given the imprecision and frequent absence of such cues, it is not surprising that the lexical status of one part of a speech stream plays an important role in determining where the beginning of the next word must be.

The long and short of word identification. One perverse feature of speech is the fact that it is not always possible to identify a word unambiguously until one has heard the word after it. Consider, for example, the word "tar". If we are listening to an utterance and have gotten just to the /r/ in "The man saw the tar box", we do not have enough information to say unequivocally that the word "tar" will not turn out to be "target" or "tarnished" or one of several other possibilities. It is only after

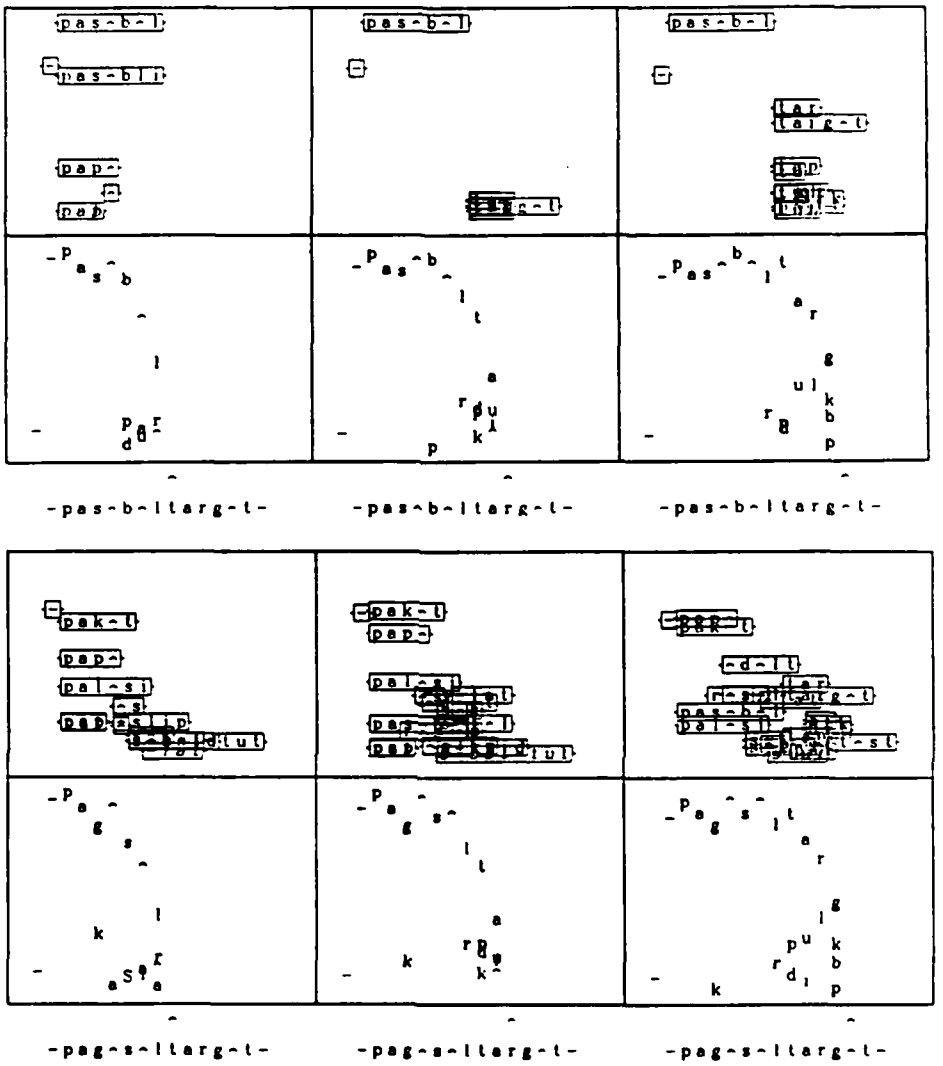


Figure 33. State of the Trace at several points during the processing of "possible target" and "pagusle target".

more time has passed, and we have perceived either a silence or enough of the next word to rule out any of the continuations of /tar/, that we can decide we have heard the word "tar". This situation, as it arises in TRACE with the simple utterance /tarbaks/ is illustrated in Figure 34. It is only well after the phoneme /b/ has registered at the phoneme level that the correct word gains the upper hand.

With longer words the situation is different. As we have already seen in another example, by the time the end of a longer word is reached it is much more likely that only one word candidate will remain. Indeed, with longer words it is often possible to have enough information to identify the word unambiguously before the end of the word. An illustration of this situation is provided by a simulation using the utterance /g^tarbaks/. By the time the /r/ has registered, "guitar" has gained the upper hand, and can be unambiguously identified without further ado.

Recently, an experiment by Grosjean (1980) has demonstrated these same effects empirically. Grosjean presented subjects with long or short words followed by a second word and measured how much of the word and its successor the subject needed to hear to identify the target. As expected, more of the next word was required to hear a short word than to hear a longer one.

In summary, then, it appears that TRACE II provides an account of two very different findings which indicate that the identification of the beginnings and endings of words depends on such factors as the length of the word and the lexical status of preceeding and following words.

General Discussion

Summary of TRACE's Successes

We feel that TRACE does a very good job of handling some of the computational challenges facing models of speech perception. At the phoneme level it has several desirable properties:

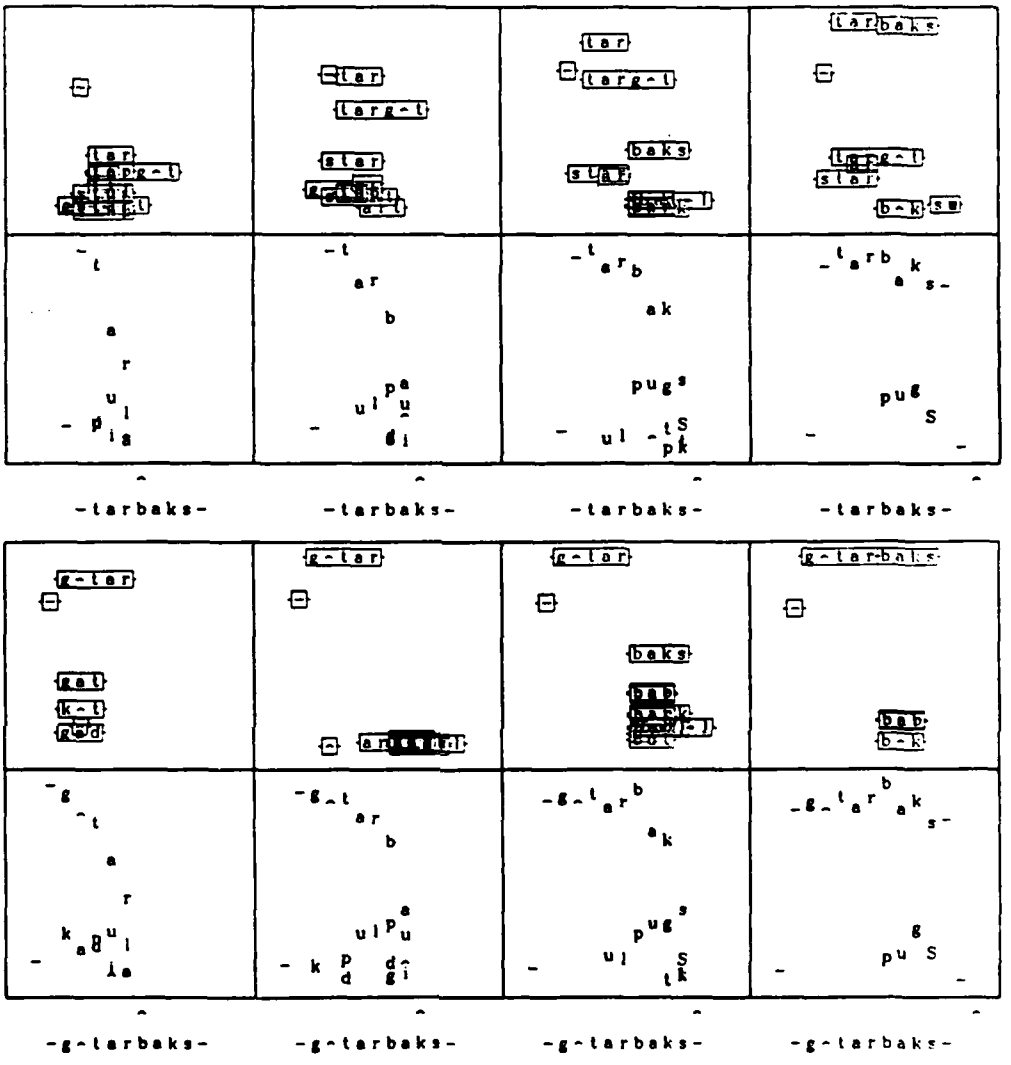


Figure 34. State of the Trace at several points in processing "tar box" and "guitar box".

1) It copes with the fact that there are no boundaries between phonemes.

2) It does not throw away information by imposing an artificial segmentation.

3) It exploits the lawful variation in the way phonemes are produced as a function of context.

4) It integrates lexical information with information extracted from the speech stream in identifying ambiguous or distorted speech sounds.

At the word level, TRACE exhibits the following desirable features:

5) It exploits information about word identity within a short time of its availability.

6) It does not depend on the clarity of word beginnings for correct word identification, and tends to pick the best overall match to the input stream, rather than relying overly strongly on the beginnings of words.

7) Word recognition is possible without any cues in the speech stream telling the model where one word ends and the next word begins.

8) Segmentation, as well as identification, can occur based on imperfect input, as when duplicate phonemes at the boundary between words are reduced.

9) It is capable of maintaining multiple word possibilities and multiple possible parses of an input stream making these available for selection or even simultaneous incorporation by higher levels.

10) It is obvious that activations descending from syntactic and semantic levels could be easily integrated with the bottom-up activations coming from the phoneme level to facilitate word recognition.

At the same time that the model shows these computational virtues, it accounts for a number of psychological findings as well. Several psychological findings derive from the fact that humans show many of the same computational advantages over previous models of speech perception. Like TRACE human speakers exhibit the following three computational advantages over other models:

1) The exploit the lawful variability in the speech wave.

2) They use information from overlapping portions of the speech wave for phoneme identification.

3) They trade features off against each other, tending to select the phoneme that has the best overall fit to the overall feature ensemble.

4) They combine information from the speech stream and from the lexical level in reaching their decisions about the identity of phonemes.

In addition, the model accounts for a number of additional facts about human speech perception. Like TRACE, humans exhibit all of the following phenomena:

5) They show a tendency toward categorical perception.

6) They exhibit lexical influences on phoneme identification in some experiments but not in others.

7) They exhibit apparent phonotactic rule effects on phoneme identification, under some situations, but there are other situations in which these effects appear to be over-ridden.

8) They show evidence of using lexical information to identify the beginnings and endings of words.

Some of the reasons for these successes. To what does the TRACE model owe its computational and psychological successes? Some of TRACE's successes simply depend on its ability to make use of the information as it comes in. It fails to show context effects only when a response must be made, or can be made with high accuracy, before contextual information is available.

There are several other reasons for TRACE's success. One, we think, is the use of continuous activation and competition processes in place of discrete decisive processes such as segmentation and identification. Activation and competition are matters of degree, and protect TRACE form catastrophic committment in marginal cases, and they provide a natural means for combining many different sources of information.

Part of the success of TRACE is specifically due to the use of competitive inhibitory interactions instead of bottom-up (or top-down) inhibition. Competition allows the model to select the best interpretation available, settling for an imperfect one when no better one is available, but overriding poor ones when a good one is at hand. These and other virtues of competitive inhibition have been noted before (e.g, Grossberg, 1973; Levin, 1976; Feldman and Ballard, 1982) in other contexts. Their usefulness here attests to the general utility of the competitive inhibition mechanism.

The elimination of between-level inhibition from the interactive activation mechanism puts us in a very nice position, with respect to one general critique of interactive activation models. It is often said that activation models are too unconstrained and too flexible to be anything more than a language for conveniently describing information processing. We are now in a position to suggest that a restricted version of the framework is sufficient. Interactive activation models could exploit both excitatory and inhibitory connections both between and within levels, but in the interactive activation model of word perception, only inhibitory interactions were allowed within a level. In TRACE, we have gone even further, allowing only excitatory connections between levels and only inhibitory connections within levels. From our experience, it appears that models which adhere to these constraints work as well or better than members of the more general class that do not. We hasten to add that we have no proof that this it true. We have, however, no reason to feel that we could improve the performance of our model by allowing either between level inhibitory interactions, or within level excitation.

Other aspects of the successes of the TRACE model depend on its use of feedback from higher to lower levels. Feedback plays a central role in the accounts of categorical perception, lexical effects on phoneme identification, and and "phonotactic rule" effects.

We do not claim that any of these phenomena, taken individually, require the assumption of a feedback mechanism. For example, consider the phenomenon of categorical perception. We use feedback from the phoneme to the feature level to drive feature patterns closer to the prototype of the phoneme they most strongly activate. This mechanism, coupled with the competition mechanism at the phoneme level, accounts for better discrimination between than within categories. However, we could account for categorical perception by suggesting that subjects do not have access to the acoustic level at all, but only to the results of the phoneme identification process. Similarly, lexical effects on

phoneme identification can be accounted for by assuming that subjects (sometimes) readout from the word level and infer the identity of phonemes from the lexical code (Morton, 1979; Marslen-Wilson and Welsh, 1978; Marslen-Wilson, 1980). In the case of "phonotactic rule" effects, other interpretations are of course available as well. One could, for example, simply suppose that subjects use knowledge of the phonotactic constraints, perhaps captured in units standing for legal phoneme pairs; and that it is the output of such units that accounts for the influence of phonotactic regularity on phoneme identification.

We know of few strong empirical reasons to prefer feedback accounts to other possibilities. However, we have two theoretical reasons for preferring to retain top-down as well as bottom-up interactions in our activation models. One reason has to do with the simplicity of the resulting decision mechanisms. Feedback allows higher-level considerations to influence the outcome of processing at lower levels in just the same way that lower-level considerations influence the outcome of processing at higher levels. The influences of lexical and other constraints on phoneme identification need not be pushed out of the theory of speech perception itself into decision processes, but are integrated directly into the perceptual process in a unified way. Given top-down as well as bottom-up processing, the decision mechanisms required for generating overt responses that reflect lexical and other contextual influences are greatly simplified; no special provision needs to be made for combining lexical and phonetic outputs in the decision mechanism.

A second reason for retaining feedback comes up when we consider the problem of learning. Although we have not discussed how learning might occur in TRACE, we have assumed that the mechanisms of speech perception are acquired through modification of connection strengths. Very roughly, in many learning schemes, connections between nodes are strengthened when two nodes tend to be activated simultaneously, at the expense of connections between nodes that tend not to be activated at the same time (c.f. Grossberg, 1978; Rosenblatt, 1962; Rumelhart and Zipser, in press).

In such schemes, however, there is a serious problem if activation is entirely bottom-up; for in that case, once a particular node has been "tuned" to respond to a particular pattern, it is difficult to retune it; it fires when its "expected" pattern is presented, and when it fires, its tendency to respond to that pattern only increases. Feedback provides a way to break this vicious cycle. If higher levels insist that a particular phoneme is present, then the node for that phoneme can become activated even if the bottom-up input would normally active some other phoneme instead; then the learning mechanism can "retune" the detector for the phoneme so that it will need to depend less on the top-down input the next time around.

In general, the use of feedback appears to place more of the intelligence required for perception and perceptual learning into the actual perceptual mechanism itself, and to make the mechanisms which exhibit this intelligence explicit. As formulated here, these mechanisms are incredibly simple; yet they appear to buy quite a lot which often gets pushed into unspecified "decision" and "guessing" processes.

The success of TRACE also depends upon its architecture, rather than the fundamental computational principles of activation and competition, or the decision to include feedback. By architecture, we mean the organization of the Trace structure into layers consisting of nodes corresponding to units at particular times. This architecture is one we decided upon only after several other kinds of architecture had failed, for it involves a massive reduplication of nodes and connections. However, it has a number of very important advantages over models which do not provide distinct tokens of the same unit to represent instances of it which might occur at different points in the Trace. There are three principle positive consequences of the Trace architecture. First, it keeps straight what occurred when in the speech stream. The model has no difficulty with misidentifying "tack" and "cat", as it would if position-invariant phoneme nodes were connected to position-invariant word nodes. Second, it permits both forward and backward interactions. Backward interactions are absolutely essential if the

model is to account for the fact that the identity of a phoneme (or a word; Warren & Sherman, 1974) can be influenced by what comes after it as well as what comes before it. Some kind of record of the past is necessary to capture these kinds of influences, as well as to provide a clear picture of the sources of the more conventional effects of preceding context, and the TRACE construct lays this out in a way that is both comprehensible and efficient. Third, the Trace structure provides an explicit mechanism which instantiates the idea that there may be a continuity between the structures involved in perceptual processing and those which provide a working memory for the results of the perceptual process. At one and the same time, the Trace is a perceptual processing system and a memory system. As a result, the model automatically accounts for the fact that coherent memory traces persist longer than incoherent ones. The coherent ones resonate through interactive (that is, bottom-up and top-down) activation, while incoherent ones fail to establish a resonance, and therefore die away more rapidly. This last point has previously been made by Grossberg (1978); his analysis of interactive activation processes in perception and memory captures the continuity of perception and memory as well as many other desirable properties of interactive activation mechanisms, though his model differs from ours in several important respects.

Some deficiencies of Trace

Although TRACE has had a number of important successes, it also has a number of equally important deficiencies. We have already mentioned the fact that it requires massive duplication of nodes and connections. Copying over and over again the connection strengths that determine which features activate which phonemes and which phonemes activate which words. As we already noted, learning in activation models (e.g., Grossberg, 1976; Rumelhart and Zipser, in press; Ackley, Hinton, and Sejnowski, in press) usually involves the retuning of connections between nodes depending on their simultaneous activation. Given the TRACE architecture, such learning would not generalize from one part of the Trace to another, and so would not be accessible for inputs arising at different

locations in the Trace. A second problem is that the model, as is, is insensitive to variation in global parameters, such as speaking rate, speaker characteristics a ' accent, and ambient acoustic characteristics. A third deficiency is that it fails to account for the fact that one presentation of a word has an effect on the perception of it a very short time later (Nusbaum and Slowiaczek, 1982). These two presentations, in the current version of the model, simply excite separate tokens for the same word in different parts of the Trace.

What these deficiencies appear to call for is a model in which there is a single stored representation of each phoneme and each word in some central representational structure. If this structure is accessed every time the word is presented, then we could account for repetition priming effects. Likewise, if there were a single central structure, learning could occur in just one set of nodes, as could dynamic retuning of feature-phoneme and phoneme-word connections to take account of changes in global parameters or speaker characteristics.

However, it remains necessary to keep straight the relative temporal location of different feature, phoneme and word activations. Thus it will not do to simply abandon the Trace in favor of a single set of nodes consisting of just one copy of each phoneme and one copy of each word.

It seems that we need to have things both ways: We need a central representation that plays a role in processing every phoneme and every word and that is subject to learning, retuning and priming. We also need to keep a dynamic trace of the unfolding representation of the speech stream, so that we can continue to accommodate both left and right contextual effects. We are currently beginning to develop a model that has these properties, based on a scheme for using one parallel processing structure to program another (McClelland, 1984).

Implications for Issues in Language and Language Perception.

There are a number of issues in language and language perception which TRACE speaks to. Here we consider three questions which seem to lie close to the heart of our conception of what language perception is all about. First, what constitutes speech perception? Second, what is the representation of linguistic rules? and third, is there anything unique or special about speech perception?

What does it mean, to perceive speech? At a number of points in this article, we have alluded to ways in which our conception of perception differs from the usage of other authors. Such concepts as perception are inherently tied to theory, and only derive their meaning with respect to particular theoretical constructs. Where does TRACE place us, then, with respect to the question, what is speech perception?

For one thing, TRACE blurs the distinction between perception and other aspects of cognitive processing. There is really no clear way in TRACE to say where perceptual processing ends and conceptual processes or memory begin. Nevertheless, we can ask, "What does it mean, to perceive speech?". Following Marr (who asked the same question for vision), we could say it means to form representations that capture the stimulus -- the speaker's utterance -- at several levels of description. Trace provides such a set of representations, as well as processes to construct them. On this view, then, the Trace is the percept, and interactive activation is the process of perception.

Aspects of this definition are appealing. For example, on this view, the percept is a very rich object, one that refers both to abstract, conceptual entities like words and perhaps at higher levels even meanings, as well as to more concrete entities like acoustic signals and features. Perception is not restricted to one or a subset of levels, as it is in certain models (e.g., Morton, 1979; Marslen-Wilson, 1980).

On the other hand, the definition seems overly liberal, for it appears that perceptual experience and access to the results of perceptual processing for the purposes of overt responding may not be completely unconstrained. A number of experiments, both in speech (e.g., Foss and Swinney, 1973; McNeil and Lindig, 1973) and reading (Healy, 1976; Drewnowski and Healy, 1977) suggest that under certain conditions lower levels of processing are inaccessible, or are at best accessed only with extra time or effort. On this evidence, if perception is to form representations, and if the representations are anything like those postulated in TRACE, then perception is quite independent of the experience of the perceiver and of access to the percept. Put another way, we may choose to define the Trace as the percept, but it is not the perceptual experience. This does not seem to be a very satisfactory state of affairs.

One coherent response to these arguments would be to say that the Trace is not the experience itself, but that some part or parts of it may be the *object* of perceptual experience. It seems sensible, for example, to suppose that the Percept itself consists of that part of the Trace under scrutiny by the decision mechanisms. On this view, it would not be incoherent to suppose that representations might be formed which would nevertheless be inaccessible either to experience or to overt response processes. It would be a matter separate from the analysis of the interactive-activation process itself to specify the scope and conditions of access to the Trace. In our simulations, we have assumed that the decision mechanism could be directed with equal facility to all levels, but this may turn out to be a simplifying assumption that does not apply in all cases.

How are rules represented? It is common in theories of language to assume without discussion that linguistic rules are represented *as such* in the mind of the perceiver, and that perception is guided primarily by consultation of such rules. However, there are a number of difficulties associated with this view. First, it does not explain how exceptions are handled; it would seem that for every excep-

tion, there would have to be a special rule that takes precedence over the more general formulation. Second, it does not explain aspects of rule acquisition by children learning language, particularly the fact that rules appear to be acquired, at least to a large extent, on a word by word basis; acquisition is marked by a gradual spread of the rule from one lexical item or set of lexical items to others. Third, it does not explain how rules come into existence; as with acquisition, it appears that rules spread gradually over the lexicon. It is difficult to reconcile several of these findings with traditional rule-based accounts of language knowledge and language processing.

Models like TRACE and the interactive activation model of word recognition take a very different perspective on the issue of linguistic rules. They are not represented as such, but rather they are built into the perceptual system via the excitatory and inhibitory connections needed for processing the particular items which embody these rules. Such a mechanism appears to avoid the problem of exceptions without difficulty, and to hold out the hope of accounting for the observation that rule acquisition and rule change are strongly tied to particular items which embody the rules.

What's special about speech?

We close by raising a question that often comes up in discussions of the mechanisms of speech perception. Is speech special? If so, in what ways? It has been argued that speech is special because of the distinctive phenomenon of categorical perception; because of the encodedness of information about one phoneme in those portions of the speech stream that are generally thought to represent other phonemes; because the information in the speech stream that indicates the presence of a particular phoneme appears not to be invariant at any obvious physical level; because of the lack of segment boundaries, and for a variety of other reasons.

Over the last several years, a number of empirical arguments have been put forward that suggest that perhaps speech may not be so special, or at least, not unique. A number of investigators have reported categorical perception in other modalities; computational approaches to problems in vision have made clear that information that must be extracted from visual displays is often complexly encoded with other information; the lack of clear boundaries between perceptual units in vision is notorious; and in reading, we higher-level perceptual objects (that is, words) appear to be defined not in terms of any visual invariants but in terms of invariant sequences of the abstract letter units they contain. Thus, the psychological phenomena that characterize the mechanisms of human speech perception, and the computational problems that must be met by any mechanism of speech perception, are not, in general, unique to speech. To be sure, the particular constellation of problems that must be solved in speech perception is different than the constellation of problems faced in any other particular case, but most of the the individual problems themselves do not appear to have analogs in other domains.

We therefore prefer to view speech as an excellent testbed for the development of an understanding of mechanisms which might turn out to have considerably broader application. Speech is special to us, since it so richly captures the multiplicity of the sources of constraint which must be exploited in perceptual processing, and because it so clearly indicates the powerful influences of the mechanisms of perception on the constructed perceptual representation. We see the TRACE model as an example of a large class of models, variously called interactive activation models, massively parallel models, or one of a variety of other names, which hold out great promise to provide a deeper understanding of the mechanisms generally used in perception.

References

Ackley, D., Hinton, G., and Sejnowski, T. (in press). Boltzmann Machines: Constraint Satisfaction networks that learn, *Cognitive Science*.

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

Anderson, J. A. (1977). Neural models with cognitive implications. In D. LaBerge & S. J. Samuels (Eds), *Basic processes in reading: Perception and comprehension*. Hillsdale, N.J.: Erlbaum Associates.

Anderson, J.A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84*, 413-451.

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

Cole, R. A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics, 13*, 153-156.

Delattre, P., Liberman, A.M., & Cooper, F.S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27*, 769-773.

Denes, P. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America, 27*, 761-764.

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

Drewnowski. A., and Healy, A. (1977). Detection Errors on *the* and *and:* Evidence for reading units larger than the word. *Memory and Congition, 5*, 636-647.

Elman, J.L., & McClelland, J.L. (1984). The Interactive Activation Model of Speech Perception. In Norman Lass (Ed.), *Language and Speech*. New York: Academic Press, pp. 337-374.

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

Fant, G. (1973). Stops in CV-syllables. In G. Fant (Ed.), *Speech Sounds and Features*. Cambridge, MA: MIT Press, pp. 110-139.

Feldman, J. A. & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 6*, 205-254.

Foss, D. J., & Blank, M. A. (1980). Identifying the speech codes. *Cognitive Psychology, 12*, 1-31.

Foss., D.J., & Swinney, D.A. (1973). On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior, 12,* 246-257.

Fox, R. (1982). Unpublished manuscript, Vanderbilt University.

Fujimura, O., & Lovins, J. B. (1982). Syllables as concatenative phonetic units. In A. Bell and J. B. Hooper (Eds.), *Syllables and segments.* Amsterdam: North-Holland, pp. 107-120.

Fujisaki, H., & Kawashima, T. (1968). The influence of various factors on the identification and discrimination of synthetic vowel sounds. Paper presented at the Sixth International Congress on Acoustics, Tokto, Japan.

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

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics,* 1980, *28,* 267-283.

Grossberg, S. (1973). Contour enhancement, short-term memory, and constancies in reverbrating neural networks. *Studies in Applied Mathematics, 52,* 217-257.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics, 23,* 121-134.

Grossberg, S. (1978). A theory of visual coding, memory, and development. In Leeuwenberg, E. L. J. & Buffart, H. F. J. M. (Eds.), *Formal theories of visual perception.* New York: John Wiley and Sons.

Grossberg, S. (1980). How does the brain build a cognitive code? *Psychological Review, 87,* 1-51.

Healy, A.F. (1976). Detection errors on the word *the*: evidence for reading units larger than letters. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 235-242.

Hinton, G. E. (1981). A parallel comutation that assigns canonical object-based frames of reference. In *Proc. IJCAI-7,* Vancouver, B. C.

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

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America, 72,* 379-389.

Kopec, G. E. (1984). Voiceless stop consonant identification using LPC spectra. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* San Diego, Pp. 42.1.1-42.1.4.

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

Levin. J. A. (1976). Proteus: An activation framework for cognitive process models (ISI/WP-2).

Marina del Rey, Calif.: Information Sciences Institute.

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

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

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

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

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

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

Massaro, D. W., & Cohen, M. M. (1983). Phonological constraints in speech perception. *Perception & Psychophysics, 34*, 338-348.

Massaro, D. W., & Oden, G. C. (1980a). 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, pp. 129-165.

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

McClelland, J. L. (in press). Putting knowledge in its place: A scheme for programming parallel processing structures on the fly. *Cognitive Science*.

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

McNeil, D., & Lindig, K. (1973). The perceptual reality of phonemes, syllables, words, and sentences. *Journal of Verbal Learning and Verbal Behavior, 12*, 419-430.

Miller, J. L. (1981). 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, pp. 39-74.

Miller, J. L., & Eimas, P. D. (1977). 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, 61*, 835-845.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*, 165-178.

Morton, J. (1979). Word recognition. In J. Morton and J. C. Marshall (Eds.), *Psycholinguistics 2:*

Structures and Processes. Cambridge, Mass.: M.I.T. Press, pp. 107-156.

Nusbaum, H.C., & Slowiaczek, L.M. (1982). An activation model of auditory word recognition. *Research on Speech Perception, Progress Report No. 8*, Department of Psychology, Indiana University. 289-305.

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

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

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

Repp, B.H. (1983). Phonetic and auditory trading relations between acoustic cues in speech perception: Further results. *Haskins Status Reports on Speech Research, SR-73*, 121-139.

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

Rosenblatt, F. (1962). *Principles of neurodynamics*. New York: Spartan Books, 1962.

Rumelhart, David E. (1977). Toward an interactive model of reading. In S. Dornic (Ed.), *Attention and performance VI*. Hillsdale, N.J.: Erlbaum.

Rumelhart, D. E., & J. L. McClelland. (1981). Interactive processing through spreading activation In C. Perfetti & A. Lesgold (Eds.), *Interactive processes in reading*, Hillsdale N.J.: Erlbaum.

Rumelhart, D. E., & McClelland, J. L. (1982). 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, 89*, 60-84.

Rumelhart, D. E. and Zipser, D. (in press). Competitive Learning. *Cognitive Science*.

Samuel, A.G. (1977). The effect of discrimination training on speech perception: Noncategorical perception. *Perception & Psychophysics, 22*, 321-330.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodoloy. *Journal of Experimental Psychology: General, 110*, 474-494.

Studdert-Kennedy, M., Liberman, A.M, Harris, K.S., Cooper, F.S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review, 77*, 234-249.

Swinney, D. A. (1982). The structure and time-course of information interaction during speech comprehension: Lexical segmentation, access, and interpretation. In J. Mehler, E.C.T. Walker, & M. Garret (Eds.), *Perspectives on mental representation*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Thompson, H. (1984). Word recognition: a Paradigm case in computational (psycho-)linguistics. Proceedings of the Sixth Annual Meeting of the Cognitive Science Society, Boulder, CO.

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

END

FILMED

1-85

DTIC