CHAPTER 15

Interactive Processes in Speech Perception: The TRACE Model

J. L. McCLELLAND and J. L. ELMAN

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. First, there are the acoustic cues to the identity of the /g/ itself. Second, the other phonemes in the same word provide another source of cues, for if we know the rest of the phonemes in this word, there are only a few phonemes that can form a word with them. Third, the semantic and syntactic context further constrain the possible words that might occur, and thus limit still further the possible interpretation of the first phoneme in *gift*.

There is ample evidence that all of these different sources of information are used in recognizing words and the phonemes they contain. Indeed, as R. A. Cole and Rudnicky (1983) have recently noted, these basic facts were described in early experiments by Bagley (1900) over 80 years ago. Cole and Rudnicky point out that recent work (which we consider in detail below) has added clarity and detail to these basic findings but has not lead to a theoretical synthesis that provides a satisfactory account of these and many other basic aspects of speech perception.

In this chapter, we describe a model that grew out of the view that the interactive activation processes that can be implemented in PDP

This chapter is a condensed version of the article "The TRACE Model of Speech Perception" by J. L. McClelland and J. L. Elman, which appeared in *Cognitive Psychology*, 1986, *18*, 1-86.

models provide a natural way to capture the integration of multiple sources of information in speech perception. This view was based on the earlier success of the interactive activation model of word perception (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982) in accounting for integration of multiple sources of information in recognizing letters in words.

In attempting to apply the ideas embodied in the interactive activation model of word perception to speech, it soon became apparent that speech provided many challenges. The model we have come up with, the TRACE model, is a response to many of these challenges and demonstrates how they can be met within the PDP framework. After we developed the model, we discovered many aspects of its behavior that are consistent with facts about speech. Thus, we were gratified to discover that the search for a mechanism that was sufficient to meet many of the challenges also lead to a model that provided quite close accounts of a number of basic aspects of the literature on speech perception.

In what follows, we begin by reviewing several facts about speech that played a role in shaping the specific assumptions embodied in TRACE. We then describe the structure of the TRACE model and the salient features of the two versions we have developed to handle different aspects of the simulations. Following this, we describe how the model accounts for a considerable body of psychological data and meets some of the computational challenges facing mechanisms of speech perception. The discussion section considers some reasons for the success of the model, explains its limitations, and indicates how we plan to overcome these in future work.

SOME IMPORTANT FACTS ABOUT SPEECH

Our intention here is not to provide an extensive survey of the nature of speech, but rather to point to several fundamental aspects of speech that have played important roles in the development of the TRACE model. A very useful discussion of several of these points is available in Klatt (1980).

Temporal Nature of the Speech Stimulus

It does not, of course, take a scientist to observe one fundamental characteristic of speech: It is a signal that is extended in time. This differentiates speech perception from most other perceptual applications

of PDP models, which have generally been concerned with visual stimuli.

The sequential nature of speech poses problems for the modeling of contextual influences, in that to account for context effects, it is necessary to keep a record of the context. It would be a simple matter to process speech if each successive portion of the speech input were processed independently of all of the others, but, in fact, this is clearly not the case. The presence of context effects in speech perception requires a mechanism that keeps some record of that context, in a form that allows it to influence the interpretation of subsequent input.

Left and Right Context Effects

A further point, and one that has been much neglected in certain models, is that it is not only prior context, but also subsequent context, that influences perception. (This and related points have recently been made by Grosjean & Gee, 1984; Salasoo & Pisoni, 1985; and Thompson, 1984). For example, Ganong (1980) reported that the identification of a syllable-initial speech sound that was constructed to be between /g/ and /k/ was influenced by whether the rest of the syllable was /ɪs/ (as in *kiss*) or /ɪft/ (as in *gift*). Such *right context effects* (Thompson, 1984) indicate that the perception of what comes in now both influences and is influenced by the perception of what comes in later. This fact suggests that the record of what has already been presented cannot be a static representation but should remain in a malleable form, subject to alteration as a result of influences arising from subsequent input.

Lack of Boundaries and Temporal Overlap

A third fundamental point about speech is that the cues to successive units of speech frequently overlap in time. The problem is particularly severe at the phoneme level. A glance at a schematic speech spectrogram (Figure 1) clearly illustrates this problem. There are no separable packets of information in the spectrogram like the separate feature bundles that make up letters in printed words.

Because of the overlap of successive phonemes, it is difficult, and we believe counterproductive, to try to divide the speech stream up into separate phoneme units in advance of identifying the units. A number of other researchers (e.g., Fowler, 1984; Klatt, 1980) have made much

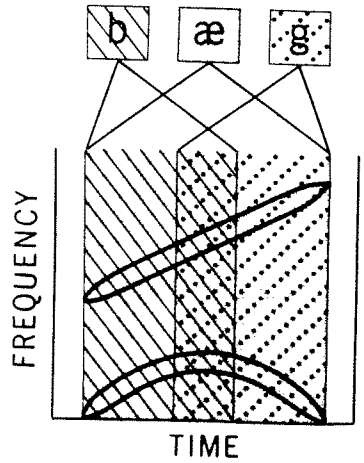

FIGURE 1. A schematic spectrogram for the syllable *bag*, indicating the overlap of the information specifying the different phonemes. (From "The Grammars of Speech and Language" by A. M. Liberman, 1970, *Cognitive Psychology*, *1*, p. 309. Copyright 1970 by Academic Press, Inc. Reprinted by permission.)

the same point. A superior approach seems to be to allow the phoneme identification process to examine the speech stream for characteristic patterns, without first segmenting the steam into separate units.

The problem of overlap is less severe for words than for phonemes, but it does not go away completely. In rapid speech, words run into each other, and there are no pauses between words. To be sure, there are often cues that signal the locations of boundaries between words—stop consonants are generally aspirated at the beginnings of stressed words in English, and word initial vowels are generally preceeded by glottal stops, for example. These cues have been studied by a number of investigators, particularly Lehiste (e.g., Lehiste, 1960, 1964) and Nakatani and collaborators. Nakatani and Dukes (1977) demonstrated that perceivers exploit some of these cues, but found that certain utterances do not provide sufficient cues to word boundaries to permit reliable perception of the intended utterance. Speech errors often involve errors of word segmentation (Bond & Garnes, 1980), and certain segmentation decisions are easily influenced by contextual factors (R. A. Cole & Jakimik, 1980). Thus, it is clear that word recognition cannot count on an accurate segmentation of the phoneme stream into separate word units, and in many cases such a segmentation would perforce exclude from one of the words a shared segment that is doing double duty in each of two successive words.

Context Sensitivity of Cues

A fourth major fact about speech is that the cues for a particular unit vary considerably with the context in which they occur. For example, the transition of the second formant carries a great deal of information about the identity of the stop consonant /b/ in Figure 1, but that formant would look quite different had the syllable been *big* or *bog* instead of *bag*. Thus, the context in which a phoneme occurs restructures the cues to the identity of that phoneme (Liberman, 1970).

Not only are the cues for each phoneme dramatically affected by preceding and following context, they are also altered by more global factors such as rate of speech (J. L. Miller, 1981), by morphological and prosodic factors such as position in the word and in the stress contour of the utterance, and by characteristics of the speaker such as size and shape of the vocal tract, fundamental frequency of the speaking voice, and dialectical variations (see Klatt, 1980, and Repp & Liberman, 1984, for discussions).

A number of different approaches to the problem have been tried by different investigators. One approach is to try to find relatively invariant—generally relational—features (e.g., Stevens & Blumstein, 1981). Another approach has been to redefine the unit so that it encompasses the context, and therefore becomes more invariant (Fujimura & Lovins, 1982; Klatt, 1980; Wickelgren, 1969). While these are both sensible and useful approaches, the first has not yet succeeded in establishing a sufficiently invariant set of cues, and the second may alleviate but does not eliminate the problem: Even units such as demisyllables (Fujimura & Lovins, 1982), context-sensitive allophones (Wickelgren, 1969), or even whole words (Klatt, 1980) are still influenced by context. We have chosen to focus instead on a third possibility: that the perceptual system uses information from the context in which an utterance occurs to alter connections dynamically, thereby effectively allowing the context to retune the perceptual mechanism in the course of processing.

Noise and Indeterminacy in the Speech Signal

To compound all the problems alluded to above, there is the additional fact that speech is often perceived under less than ideal circumstances. While a slow and careful speaker in a quiet room may produce sufficient cues to allow correct perception of all of the phonemes in an utterance without the aid of lexical or other higher-level

constraints, these conditions do not always obtain. People can correctly perceive speech under quite impoverished conditions if it is semantically coherent and syntactically well-formed (G. A. Miller, Heise, & Lichten, 1951). This means that the speech mechanisms must be able to function, even with a highly degraded stimulus. In particular, as Grosjean and Gee (1984), Norris (1982), and Thompson (1984) have pointed out, the mechanisms of speech perception cannot count on accurate information about any part of a word. As we shall see, this fact poses a serious problem for one of the best current psychological models of the process of spoken word recognition, the COHORT model of Marslen-Wilson and Welsh (1978).

Many of the characteristics that we have reviewed differentiate speech from print—at least, from very high quality print on white paper—but it would be a mistake to think that similar problems are not encountered in other domains. Certainly, the sequential nature of spoken input sets speech apart from vision, in which there can be some degree of simultaneity of input. However, the problems of ill-defined boundaries, context sensitivity of cues, and noise and indeterminacy are central problems in vision just as much as they are in speech (cf. Ballard, Hinton, & Sejnowski, 1983; Barrow & Tenenbaum, 1978; Marr, 1982). Thus, though the model we present here is focused on speech perception, we would hope that the ways in which it deals with the challenges posed by the speech signal will be applicable in other domains.

The Importance of the Right Architecture

All of the considerations listed above played an important role in the formulation of the TRACE model. The model is an instance of a PDP model, but it is by no means the only instance of such a model that we have considered or that could be considered. Other formulations we considered simply did not appear to offer a satisfactory framework for dealing with these central aspects of speech (see Elman & McClelland, 1984, for discussion). Thus, the TRACE model hinges on the particular processing architecture it proposes for speech perception as well as on the PDP mechanisms that implement the interactive activation processes that occur within it.

Sources of TRACE's architecture. The inspiration for the architecture of TRACE goes back to the HEARSAY speech understanding system (Erman & Lesser, 1980; Reddy, Erman, Fennell, & Neely, 1973). HEARSAY introduced the notion of a Blackboard, a structure similar

to the Trace in the TRACE model. The main difference is that the Trace is a dynamic processing structure that is self-updating, while the Blackboard in HEARSAY was a passive data structure through which autonomous processes shared information. The architecture of TRACE also bears a resemblance to the *neural spectrogram* proposed by Crowder (1978; 1981) to account for interference effects between successive items in short-term memory.

THE TRACE MODEL

The TRACE model consists primarily of a very large number of units, organized into three levels, the *feature*, *phoneme*, and *word* levels. Each unit stands for an hypothesis about a particular perceptual object—feature, phoneme, or word—occurring at a particular point in time defined relative to the beginning of the utterance. Thus, the TRACE model uses local representation.

A small subset of the units in TRACE II, the version of the model with which we will be mostly concerned, is illustrated in Figures 2, 3, and 4. Each of the three figures replicates the same set of units, illustrating a different property of the model in each case. In the figures, each rectangle corresponds to a separate processing unit. The labels on the units and along the side indicate the spoken object (feature, phoneme, or word) for which each unit stands. The left and right edges of each rectangle indicate the portion of the input the unit spans.

At the feature level, there are several banks of feature detectors, one for each of several dimensions of speech sounds. Each bank is replicated for each of several successive moments in time, or time slices. At the phoneme level, there are detectors for each of the phonemes. There is one copy of each phoneme detector centered over every three time-slices. Each unit spans six time slices, so units with adjacent centers span overlapping ranges of slices. At the word level, there are detectors for each word. There is one copy of each word detector centered over every three feature slices. Here, each detector spans a stretch of feature slices corresponding to the entire length of the word. Again, then, units with adjacent centers span overlapping ranges of slices.

Input to the model, in the form of a pattern of activation to be applied to the units at the feature level, is presented sequentially to the feature-level units in successive slices, as it would be if it were a real stream of speech. Mock-speech inputs on the three illustrated dimensions for the phrase *tea cup* (/tik^p/) are shown in Figure 2. At any instant, input is arriving only at the units in one slice at the feature

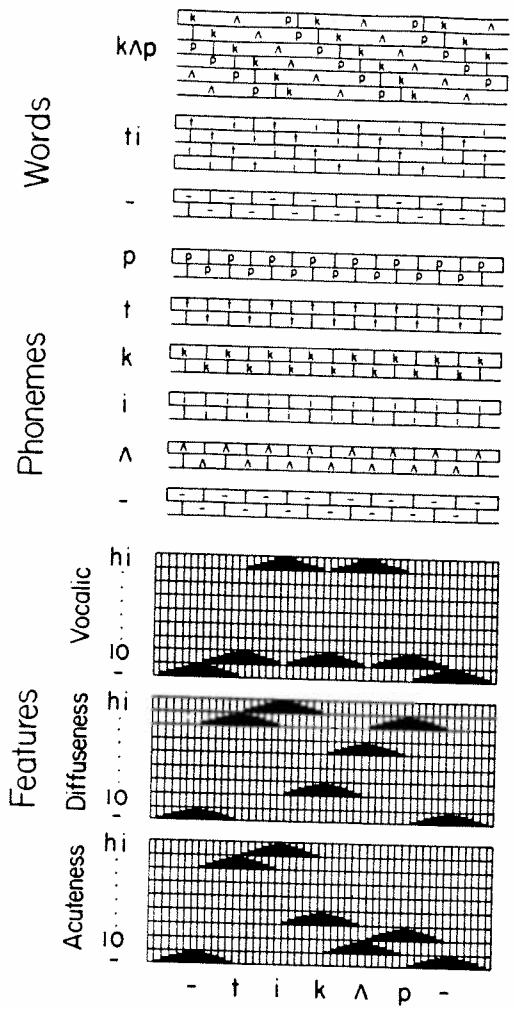

FIGURE 2. A subset of the units in TRACE II. Each rectangle represents a different unit. The labels indicate the item for which the unit stands, and the horizontal edges of the rectangle indicate the portion of the Trace spanned by each unit. The input feature specifications for the phrase *tea cup*, preceded and followed by silence, are indicated for the three illustrated dimensions by the blackening of the corresponding feature units.

level. In terms of the display in Figure 2, then, we can visualize the input being applied to successive slices of the network at successive moments in time. However, it is important to remember that all the units are continually involved in processing, and processing of the input

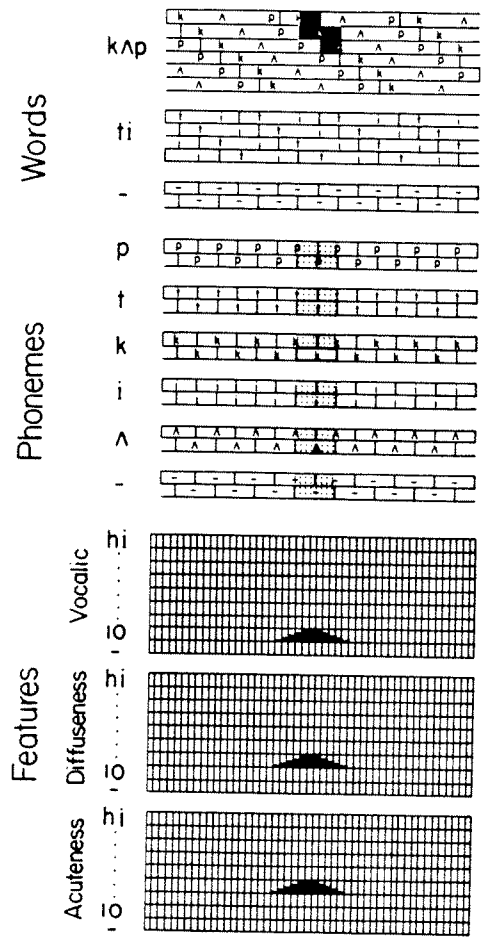

FIGURE 3. The connections of the unit for the phoneme /k/, centered over Time-Slice 24. The rectangle for this unit is highlighted with a bold outline. The /k/ unit has mutually excitatory connections to all the word- and feature-level units colored either partly or wholly in black. The more coloring on a unit's rectangle, the greater the strength of the connection. The /k/ unit has mutually inhibitory connections to all of the phoneme-level units colored partly or wholly in grey. Again, the relative amount of inhibition is indicated by the extent of the coloring of the unit; it is directly proportional to the extent of the temporal overlap of the units.

arriving at one time is just beginning as the input is moved along to the next time slice.

The entire network of units is called *the Trace*, because the pattern of activation left by a spoken input is a trace of the analysis of the input at

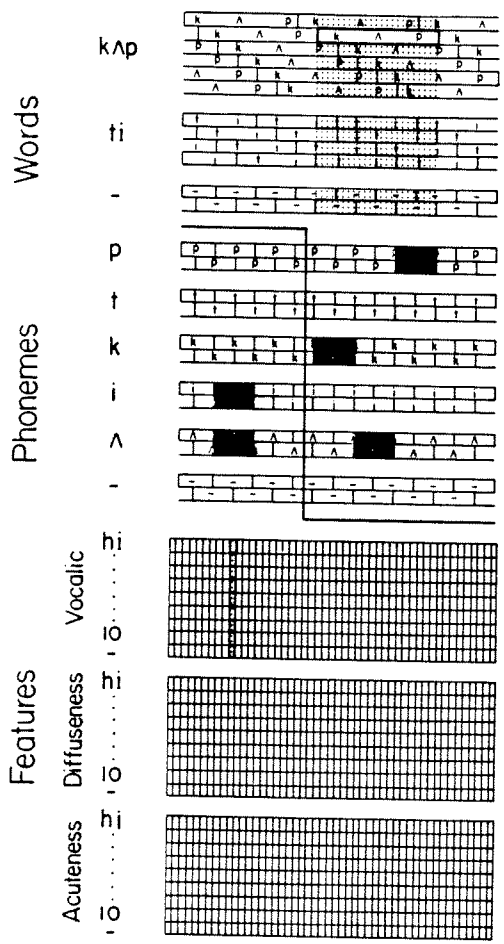

FIGURE 4. The connections of the highlighted unit for the high value on the vocalic feature dimension in Time-Slice 9 and for the highlighted unit for the word /kˆp/ starting in Slice 24. Excitatory connections are represented in black, inhibitory connections in gray, as in Figure 3.

each of the three processing levels. This trace is unlike many traces, though, in that it is active since it consists of activations of processing elements, and these processing elements continue to interact as time goes on. The distinction between perception and (primary) memory is completely blurred since the percept is unfolding in the same structures that serve as working memory, and perceptual processing of older portions of the input continues even as newer portions are coming into the

system. These continuing interactions permit the model to incorporate right context effects and allow the model to account directly for certain aspects of short-term memory, such as the fact that more information can be retained for short periods of time if it hangs together to form a coherent whole.

Processing takes place through the excitatory and inhibitory interactions of the units in the Trace. Units on different levels that are mutually consistent have mutually excitatory connections, while units on the same level that are inconsistent have mutually inhibitory connections. All connections are bidirectional. Thus, the unit for the phoneme /k/ centered over Feature-Slice 24 (shown in Figure 3) has bidirectional excitatory connections to feature units that would be activated if the input contained that phoneme centered on Time-Slice 24. It also has bidirectional excitatory connections to all the units at the word level for words containing /k/ at Time-Slice 24. The connections of illustrative feature- and word-level units are shown in Figure 4. Units on the same level are mutually incompatible, and hence mutually inhibitory, to the extent that the input patterns they stand for would overlap with each other in time. That is to say, units on the same level inhibit each other in proportion to the extent of the overlap of their temporal spans, or windows. At the feature level, units stand for the content of only a single time slice, so they only compete with units standing for other values on the same dimension (see Figure 4). At the phoneme and word level, however, there can be different degrees of overlap, and hence of mutual inhibition. The extent of the mutual inhibition between the /k/ in Slice 24 and other phoneme-level units is indicated in Figure 3 by the amount of shading that falls over the rectangle for the other unit. Similarly, the extent of mutual inhibition between the unit for /k^p/ starting in Slice 24 and other word-level units is indicated in Figure 4.

Context-Sensitive Tuning of Phoneme Units

The connections between the feature and phoneme levels determine what pattern of activations over the feature units will most strongly activate the detector for each phoneme. To cope with the fact that the features representing each phoneme vary according to the phonemes surrounding them, the model uses multiplicative connections of the kind proposed by Hinton (1981b) and discussed in Chapters 4 and 16. These multiplicative connections essentially adjust the connections from units at the feature level to units at the phoneme level as a function of activations at the phoneme level in preceding and following time slices.

For example, when the phoneme /t/ is preceded or followed by the vowel /i/, the feature pattern corresponding to the /t/ is very different than it is when the /t/ is preceded or followed by another vowel, such as /a/. Accordingly, when the unit for /i/ in a particular slice is active, it changes the pattern of connections for units for /t/ in preceding and following slices.

TRACE I and TRACE II

In developing TRACE and in trying to test its computational and psychological adequacy, we found that we were sometimes led in rather different directions. We wanted to show that TRACE could process real speech, but to build a model that did so, it was necessary to worry about exactly what features must be extracted from the speech signal, about differences in duration of different features of different phonemes, and about how to cope with the ways in which features and feature durations vary as a function of context. Obviously, these are important problems, worthy of considerable attention. However, concern with these issues tended to obscure attention to the fundamental properties of the model and the model's ability to account for basic aspects of the psychological data obtained in many experiments.

To cope with these conflicting goals, we have developed two different versions of the model, called TRACE I and TRACE II. Both models spring from the same basic assumptions, but focused on different aspects of speech perception. TRACE I was designed to address some of the challenges posed by the task of recognizing phonemes from real speech. This version of the model is described in detail in Elman and McClelland (in press). With this version of the model, we have been able to show that the TRACE framework could indeed be used to process real speech—albeit from a single speaker uttering isolated monosyllables at this point. We have also demonstrated the efficacy of the idea of using multiplicative connections to adjust feature-to-phoneme connections on the basis of activations produced by surrounding context.

The second version of the model, TRACE II, will be the main focus of this chapter. We developed this version of the model to account for lexical influences on phoneme perception and for what is known about on-line recognition of words, though we will use it to illustrate how certain other aspects of phoneme perception fall out of the TRACE framework. This version of the model is actually a simplified version of TRACE I. Most importantly, we eliminated the connection-strength adjustment facility, and we replaced the real speech inputs to TRACE I

with mock speech. This mock-speech input consisted of overlapping but contextually invariant specifications of the features of successive phonemes. Thus, TRACE II sidesteps many of the issues addressed by TRACE I, but it makes it much easier to see how the mechanism can account for a number of aspects of phoneme and word recognition. A number of further simplifying assumptions were made to facilitate examination of basic properties of the interactive activation processes taking place within the model.

Implementation Details

The material in this section is included for completeness, but the basic line of development may be followed without reading it. Readers uninterested in these details may wish to skip to the section on factors influencing phoneme identification.

Units and their dynamics. The dynamic properties of the units in TRACE are the same as those used in the interactive activation model of visual word perception; these are described in detail in Chapter 2. In brief, the model is a synchronous model, in that all the units update their activation at the same time, based on the activations computed in the previous update cycle. Each unit takes a sum of the excitatory and inhibitory influences impinging on it. Each influence is essentially the product of the output of the influencing unit and the weight on the connection between it and the receiver. If this net input is positive, it drives the activation of the unit upward in proportion to the distance left to the fixed maximum activation level; if the net input is negative, it drives the activation of the unit down in proportion to the distance left to the fixed minimum. Activations also tend to decay back to their resting activation level, which was fixed at 0 for all units. The output of a unit is 0 if the activation is less than or equal to 0; otherwise it is equal to its activation.

TRACE I. The inputs to TRACE I are sets of 15 parameter values extracted at 5 msec intervals from syllables spoken by a male native speaker of English. The bulk of the TRACE I simulations have been done with a set of CV syllables consisting of an unvoiced stop consonant (/p/, /t/, or /k/) followed by one of the vowels /a/, /i/, and /u/, as in the words *shah*, *tea*, and *who*. At the feature level, TRACE I consists of detectors for each of eight different value ranges on each of the 15 input parameters. There is a complete set of detectors for each 5 msec time slice of the input. Since there are 100 slices, the model is capable of processing 500 msec samples of speech.

There are no word-level units in TRACE I. However, there are phoneme-level units for each successive 15 msec time slice of the speech. The connections from the feature to the phoneme units were determined by using the perceptron convergence procedure (see Chapter 2) under two different conditions. First, in the invariant connections condition, a single set of connection strengths was found for each phoneme, using tokens of the phoneme spoken in all different contexts. In the context-sensitive connections condition, separate sets of connection strengths were found for each stop consonant in the context of each of the vowels.

TRACE I can be tested either using the invariant connections or using the multiplicative context-sensitive connections described above. In the latter case, the weights coming into a particular phoneme are weighted according to the relative activation of other phonemes in the surrounding context. Consider an arbitrary phoneme unit which we will designate, for now, the target unit. The strengths of the connections coming into this unit can be designated by the vector $\mathbf{w}$, where the elements of the vector are just the individual weights from each feature unit to the phoneme unit. This vector is the average over all context phonemes k of the context-specific weight vectors appropriate for the target phoneme in the context of k, where the contribution of each of these context-specific weight vectors is proportional to the exponential of the activation of phoneme k summed over the time slices adjacent to the target phoneme unit (see Elman & McClelland, 1986, in press, for further details).

TRACE II. Inputs to TRACE II are not real speech, but *mock speech* of the kind illustrated in Figure 2. The mock speech is a series of specifications for inputs to units at the feature level, one for each 25 msec time slice of the mock utterance. These specifications are generated by a simple computer program from a sequence of to-be-presented segments provided by the human user of the simulation program. The allowed segments consist of the stop consonants /b/, /p/, /d/, /t/, /g/, and /k/; the fricatives /s/ and /S/ (*sh* as in *ship*); the liquids /l/ and /r/; and the vowels /a/ (as in *pot*), /i/ (as in *beet*), /u/ (as in *boot*), and /^/ (as in *but*). /^/ is also used to represent reduced vowels such as the second vowel in *target*. There is also a "silence" segment represented by /-/. Special segments, such as a segment halfway between /b/ and /p/, can be constructed as well.

A set of seven dimensions is used in TRACE II to represent the feature-level inputs. Of course, these dimensions are intentional simplifications of the real acoustic structure of speech, in much the same way that the font used by McClelland and Rumelhart (1981) in the interactive activation model of visual word recognition was an

intentional simplification of the real structure of print. Each dimension is divided into eight value ranges. Each phoneme has a value on each dimension; the values on the vocalic, diffuseness, and acuteness dimensions for the phonemes in the utterance /tikˆp/ are shown in Figure 2. The dimensions and the values assigned to each phoneme on each dimension are indicated in Table 1. Numbers in the cells of the table indicate which value on the indicated dimension is most strongly activated by the feature pattern for the indicated phoneme. Values range from 1 (very low) to 8 (very high). The last two dimensions were altered for the categorical perception and trading relations simulations, as described below.

Values are assigned to approximate the values real phonemes would have on these dimensions and to make phonemes that fall into the same phonetic category have identical values on many of the dimensions. Thus, for example, all stop consonants are assigned the same values on the power, vocalic, and consonantal dimensions. We do not claim to have captured the details of phoneme similarity exactly. Indeed, one cannot do so in a fixed feature set because the similarities vary as a function of context. However, the feature sets do have the property that the feature pattern for one phoneme is more similar to the feature pattern for other phonemes in the same phonetic category (stop, fricative, liquid, or vowel) than it is to the patterns for phonemes

TABLE 1

PHONEME FEATURE VALUES USED IN TRACE II

PHONEME	POW	VOC	DIF	ACU	CON	VOI	BUR
p	4	1	7	2	8	1	8
b	4	1	7	2	8	7	7
t	4	1	7	7	8	1	6
d	4	1	7	7	8	7	5
k	4	1	2	3	8	1	4
g	4	1	2	3	8	7	3
s	6	4	7	8	5	1	-
S	6	4	6	4	5	1	-
r	7	7	1	2	3	8	-
l	7	7	2	4	3	8	-
a	8	8	2	1	1	8	-
i	8	8	8	8	1	8	-
u	8	8	6	2	1	8	-
ˆ	7	8	5	1	1	8	-

POW = power, VOC = vocalicness, DIF = diffuseness, ACU = acuteness, CON = consonantal, VOI = voicing, BUR = burst amplitude. Only the stops have values on this last dimension.

in other categories. Among the stops, those phonemes sharing place of articulation or voicing are more similar than those sharing neither attribute.

The feature specification of each phoneme in the input stream extends over 11 time slices of the input. The strength of the pattern grows to a peak at the sixth slice and falls off again, as illustrated in Figure 2. Peaks of successive phonemes are separated by six slices. Thus, specifications of successive phonemes overlap, as they do in real speech (Fowler, 1984; Liberman, 1970).

Generally, there are no cues in the speech stream to word boundaries—the feature specification for the last phoneme of one word overlap with the first phoneme of the next in just the same way feature specifications of adjacent phonemes overlap within words. However, entire utterances presented to the model for processing—be they individual syllables, words, or strings of words—are preceded and followed by silence. Silence is not simply the absence of any input; rather, it is a pattern of feature values, just like the phonemes. Thus, a ninth value on each of the seven dimensions is associated with silence. These values are actually outside the range of values that occurred in the phonemes themselves so that the features of silence are completely uncorrelated with the features of any of the phonemes.

TRACE II contains a unit for each of the nine values on each of the seven dimensions, in each time slice of the Trace. At the phoneme level, each Trace contains a detector for each of the 15 phonemes and a detector for the presence of silence. The silence detectors are treated like all other phoneme detectors. Each member of the set of detectors for a particular phoneme is centered over a different time-slice at the feature level, and the centers are spaced three time-slices apart. The unit centered over a particular slice receives excitatory input from feature units in a range of 11 slices, extending both forward and backward from the slice in which the phoneme unit is located. It also sends excitatory feedback down to the same feature units in the same range of slices.

The connection strengths between the feature-level units and a particular phoneme-level unit exactly match the feature pattern the phoneme is given in its input specification. Thus, as illustrated in Figure 3, the strengths of the connections between the unit for /k/ centered over Time-Slice 24 and the units at the feature level are exactly proportional to the pattern of input to the feature level produced by an input specification containing the features of /k/ centered in the same time slice.

TRACE II also contains detectors for the 211 words found in a computerized phonetic word list that met all of following criteria: (a) The word consisted only of phonemes in the list above; (b) it was not an

inflection of some other word that could be made by adding *ed*, *s*, or *ing*; and (c) the word together with its *ed*, *s*, and *ing* inflections occurred with a frequency of 20 or more per million in the Kucera and Francis (1967) word count. It is not claimed that the model's lexicon is an exhaustive list of words meeting these criteria since the computerized phonetic lexicon was not complete, but it is reasonably close to this. To make specific points about the behavior of the model, detectors for the following three words not in the main list were added: *blush*, *regal*, and *sleet*. The model also has detectors at the word level for silence (/-/), which is treated like a one-phoneme word.

Presentation and processing of an utterance. Before processing of an utterance begins, the activations of all of the units are set at their resting values. At the start of processing, the input to the initial slice of feature units is applied. Activations are then updated, ending the initial time cycle. On the next time cycle, the input to the next slice of feature units is applied, and excitatory and inhibitory inputs to each unit resulting from the pattern of activation left at the end of the previous time slice are computed.

It is important to remember that the input is applied, one slice at a time, proceeding from left to right as though it were an ongoing stream of speech "writing on" the successive time slices of the Trace. The interactive activation process is occurring throughout the Trace on each time slice, even though the external input is only coming in to the feature units one slice at a time. Processing interactions can continue even after the left to right sweep through the input reaches the end of the Trace. Once this happens, 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 predetermined number of time cycles has elapsed.

Activations and overt responses. Activations of units 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 units located within a small window of adjacent slices within any level. The units in this set then constitute the set of response alternatives, designated by the identity of the item for which the unit stands (note that with several adjacent slices included in the set, several units in the alternative set may correspond to the same overt response). Word-identification responses are assumed to be based on readout from the word level, and phoneme-identification responses are assumed to be based on readout from the phoneme level.

As far as phoneme identification is concerned, then, a homogeneous mechanism is assumed to be used with both word and nonword stimuli. The decision mechanism can be asked to make a response either (a) at a critical time during processing, or (b) when a unit in the alternative set reaches a critical 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:

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

where j indexes the members of the alternative set, and $S_i = e^{ka_i}$. The exponential transformation ensures that all activations are positive and gives great weight to stronger activations; 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. At the expense of considerable realism, we have tried to keep both TRACE I and TRACE II simple by using homogeneous parameters wherever possible. The strength of the total excitation coming into a particular phoneme unit from the feature units is normalized to the same value for all phonemes, thus making each phoneme equally excitable by its own canonical pattern. Other simplifying assumptions should be noted as well. For example, there are no differences in connections or resting levels for words of different frequency. It would have been a simple matter to incorporate frequency as McClelland and Rumelhart (1981) did, and a complete model would, of course, include some account for the ubiquitous effects of word frequency. We left it out here to facilitate an examination of the many other factors that appear to influence the process of word recognition in speech perception.

Even with all the simplifications described above, TRACE II still has 10 free parameters; these are listed in Table 2. There was some trial and error in finding the set of parameters used in the reported simulations, but, in general, the qualitative behavior of the model is remarkably robust under parameter variations, and no systematic search of the space of parameters is necessary.

In all the reported simulations using TRACE II, the parameters were held at the values given in Table 2. The only exception to this occurred in the simulations of categorical perception and trading relations. Since we were not explicitly concerned with the effects of feedback to the feature level in any of the other simulations, we set the

TABLE 2

PARAMETERS OF TRACE II

Parameter	Value
Feature-Phoneme Excitation	.02
Phoneme-Word Excitation	.05
Word-Phoneme Excitation	.03
Phoneme-Feature Excitation	.00
Feature-Level Inhibition	.04
Phoneme-Level Inhibition*	.04
Word-Level Inhibition*	.03
Feature-Level Decay	.01
Phoneme-Level Decay	.03
Word-Level Decay	.05

*Per 3 time slices of overlap.

feedback from the phoneme level to the feature level to zero to speed up the simulations in all other cases. In the categorical perception and trading relations simulations this parameter was set at 0.05. Phoneme-to-feature feedback tended to slow the effective rate of decay at the feature level and to increase the effective distinctiveness of different feature patterns. Rate of decay of feature level activations and strength of phoneme-to-phoneme competition were set to 0.03 and 0.05 to compensate for these effects. No lexicon was used in the categorical perception and trading relations simulations, which is equivalent to setting the phoneme-to-word excitation parameter to zero. In TRACE I, the parameters were tuned separately to compensate for the finer time scale of that version of the model.

FACTORS INFLUENCING PHONEME IDENTIFICATION

We are ready to examine the performance of TRACE, to see how well it can account for psychological data on the process of speech perception and, to determine how well it can cope with the computational challenges posed by speech. In this section we consider the process of phoneme identification. In the next section we examine several aspects of word recognition. The sections may be read independently, in either order.

In the introduction, we motivated the approach taken in the TRACE model in general terms. In this section, we will see that the simple concepts that lead to TRACE provide the basis for a coherent and synthetic account of a large number of different kinds of findings on the

perception of phonemes. Previous models have been able to provide fairly accurate accounts of a number of these phenomena. For example, Massaro and Oden's feature integration model (Massaro, 1981; Massaro & Oden, 1980a, 1980b; Oden & Massaro, 1978) accounts in detail for a large body of data on the influences of multiple cues to phoneme identity, and the Pisoni/Fujisaki-Kawashima model of categorical perception (Fujisaki & Kawashima, 1968; Pisoni, 1973, 1975) accounts for a large body of data on the conditions under which subjects can discriminate sounds within the same phonetic category. Marslen-Wilson's COHORT model (Marslen-Wilson & Welsh, 1978) can account for the time course of certain aspects of lexical influences on phoneme identification. Recently Fowler (1984) has proposed an interesting account of the way listeners cope with coarticulatory influences on the acoustic parameters of speech sounds. Here we will show that TRACE brings these phenomena, and several others not considered by any of these other models, together into a coherent picture of the process of phoneme perception as it unfolds in time.

This section consists of four main parts. The first focuses on lexical effects on phoneme identification and the conditions under which these effects are obtained. The second part of this section focuses on the question of the role of phonotactic rules—that is, rules specifying which phonemes can occur together in English—in phoneme identification. Here, we see how TRACE mimics the apparently rule-governed behavior of human subjects, in terms of a "conspiracy" of the lexical items that instantiate the rule. The third part focuses on two aspects of phoneme identification often considered quite separately from lexical effects—namely, the contrasting phenomena of cue tradeoffs in phoneme perception and categorical perception. The simulations in the first three parts were all done using TRACE II. The fourth part describes our simulations with TRACE I, illustrating how the connection-modulation mechanisms embedded in that version of the model account for the fact that listeners appear to alter the cues they use to identify phonemes in different contexts.

Lexical Effects

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

[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).

phoneme identification reported by Ganong (1980). We start with this phenomenon because it, and the related phonemic restoration effect, were among the primary reasons why we felt that the interactive activation mechanisms provided by PDP models would be appropriate for speech perception, as well as visual word recognition and reading.

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 /l/, then /ʌ/, then /g/. Figure 5 shows phoneme- and word-level activations at several points in the unfolding of this input specification. Each panel of the figure represents a different point in time during the presentation and concomitant processing of the input. The upper portion of each panel is used to display activations at the word level; the lower panel is used for activations at the phoneme level. Each unit is represented by a rectangle labeled with the identity of the item the unit stands for. The horizontal extension of the rectangle indicates the portion of the input spanned by the unit. The vertical position of the rectangle indicates the degree of activation of the unit. In this and subsequent figures, activations of the phoneme units 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 input itself is indicated below each panel, with the successive phonemes positioned at the temporal positions of the centers of their input specifications. The "ʌ" along the x-axis represents the point in the presentation of the input stream at which the snapshot was taken.

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 phoneme level simply has not evolved to the point where it provides enough constraint to select any one particular word. It is only after the /g/ has come in that the model has information telling it whether the input is closer to *plug*, *plus*, *blush*, or *blood* (TRACE's lexicon contains no other words beginning with /plʌ/ or /blʌ/). After that point, as illustrated in the fourth panel, *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.

Factors influencing the lexical effect. There is now a reasonable body of literature on lexical effects on phoneme identification. One important property of this literature is the fact that the lexical effect is

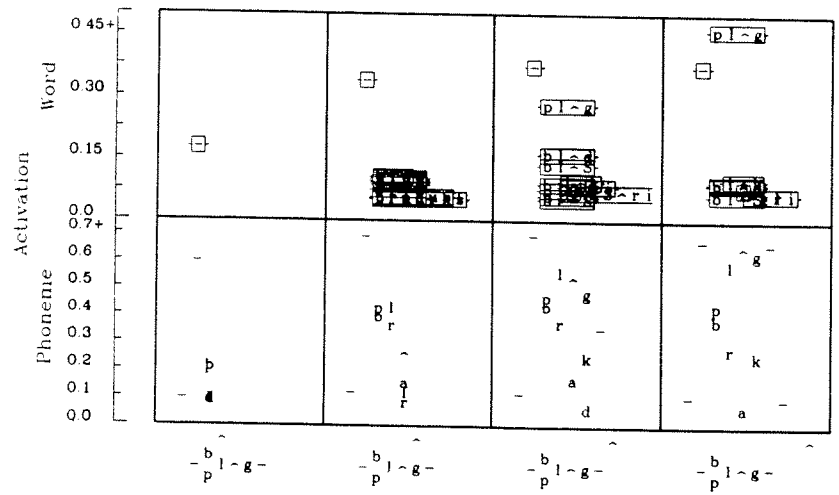

FIGURE 5. Phoneme- and word-level activations at several points in the unfolding of a segment ambiguous between /b/ and /p/, followed by /l/, /ʌ/, and /g/. See text for a full explanation.

often somewhat difficult to obtain. For example, Fox (1982, 1984) found that the lexical effect can be eliminated by time pressure. Ganong reported that the lexical effect only shows up with segments that are ambiguous; we know that in running speech, people often perceived as correctly pronounced words with deliberate errors (Marslen-Wilson & Welsh, 1978), but at the beginnings of isolated words lexical influences appear to lead to misperceptions of unambiguous tokens of phonemes. In reaction time studies, it has been observed by Foss and Blank (1980) that there is no lexical effect on the reaction time to detect word-initial phonemes.

Many of these findings have been taken as evidence against the view that top-down influences really play a role in normal perceptual processing (Foss & Gernsbacher, 1983), and only come into play in a post-perceptual stage of processing (Fox, 1982). However, we observe the same results in simulations with TRACE, where top-down influences are always at work. The reason why lexical effects do not emerge until late in processing for word-initial targets is simply that the contextual information is not available until then. The reason why lexical effects do not emerge with word-initial targets that are not ambiguous is simply that the bottom-up information is there to identify the target, long before the contextual information would be available. Simulations demonstrating the absence of lexical effects for word-initial segments

under speeded conditions or when the segment is unambiguous are described in McClelland and Elman (in press).

The crucial observations concern what happens with lexical effects on word-final segments. It is well known that lexical effects are larger later in words than they are at the beginnings of words (Marslen-Wilson & Welsh, 1978) and can be obtained in reaction time studies even with unambiguous segments (Marslen-Wilson, 1980).

TRACE produces stronger lexical effects when the target comes late in the word, simply because the context is already providing top-down support for the target when it starts to come in under these circumstances. We illustrate by comparing response strength for the phoneme /t/ in /sikr^t/ (the word *secret*) and in the nonword /g^ld^t/ (*guldut*) in Figure 6. 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 units centered 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 0.9 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/. The size of the effect Marslen-Wilson (1980) obtains is quite comparable to the effect observed in Figure 6.

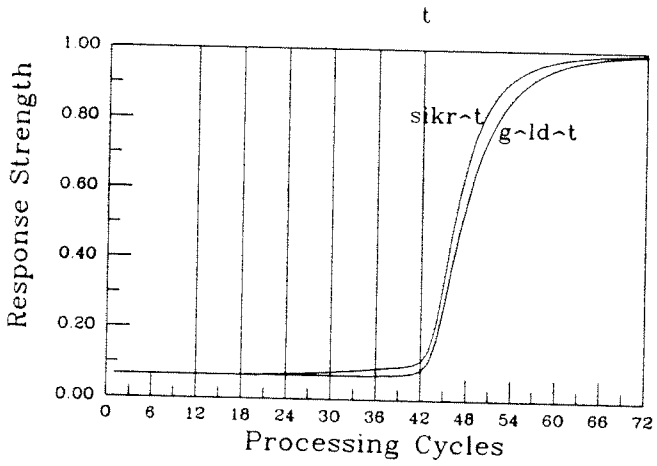

FIGURE 6. Probability of the /t/ response as a function of processing cycles, based on activation of phoneme units at Cycle 42, for the stream /sikr^t/ (*secret*) and /g^ld^t/ (*guldut*). Vertical lines indicate the peaks of the input patterns corresponding to the successive phonemes in either stream.

Are Phonotactic Rule Effects the Result of a Conspiracy?

Recently, Massaro and M. M. Cohen (1983) have reported evidence they take as support for the use of phonotactic rules in phoneme identification. In one experiment, Massaro and Cohen's stimuli consisted of phonological segments 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 TRACE 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; it would also activate *sleeve*, *sleek*, and others in a model with a fuller lexicon). 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 dominate the /r/, just as it would if /sli/ were an actual word, as shown in Figure 7.

The hypothesis that phonotactic rule effects are really based on word activations leads to a prediction: 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 (1983) 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/) was phonotactically acceptable, while the other (the /b/) was not. However, 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*).

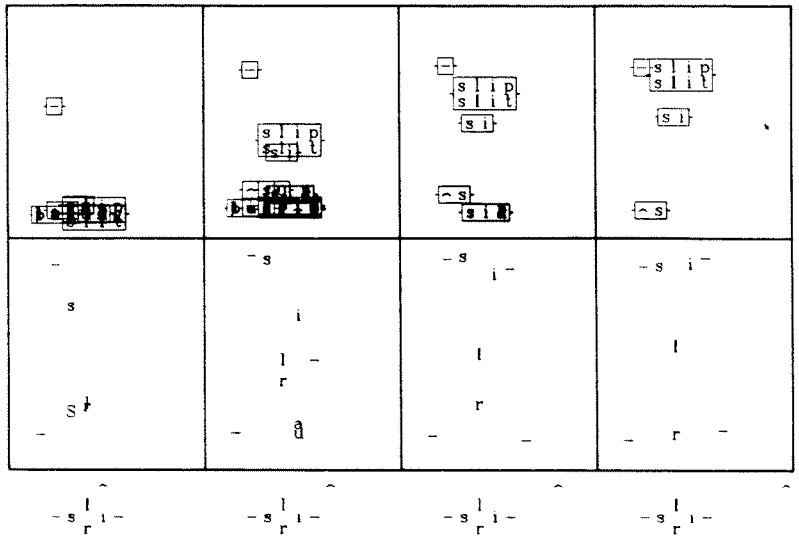

FIGURE 7. State of the Trace at several points in processing a segment ambiguous between /l/ and /r/ in the context /s_i/. The units for *sleep* (/slip/) and *sleet* (/slit/) are boxed together since they take on identical activation values.

Results of the experiment are shown in Table 3. The existence of a word identical to one of the two alternatives or differing from one of the alternatives by a single phonetic feature of one phoneme strongly influenced the subjects' choices between the two alternatives. Indeed, in the case where the phonotactically irregular alternative (*bwacelet*) was one feature away from a particular lexical item (*bracelet*), subjects tended to hear the ambiguous item in accord with the similar lexical item (that is, as a /b/) even though it was phonotactically incorrect.

TABLE 3

PERCENT CHOICE OF PHONOTACTICALLY IRREGULAR CONSONANT

Stimulus Type	Example	Percentage of Identifications as "Illegal" Phoneme (/b/)*
Legal word/illegal nonword	dwindle/bwindle	37%
Legal nonword/illegal nonword	dwiffle/bwiffle	46%
Legal nonword/illegal near-word	dwacelet/bwacelet	55%

*$F(2,34) = 26.414$, $p < .001$

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 8. 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 the 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.

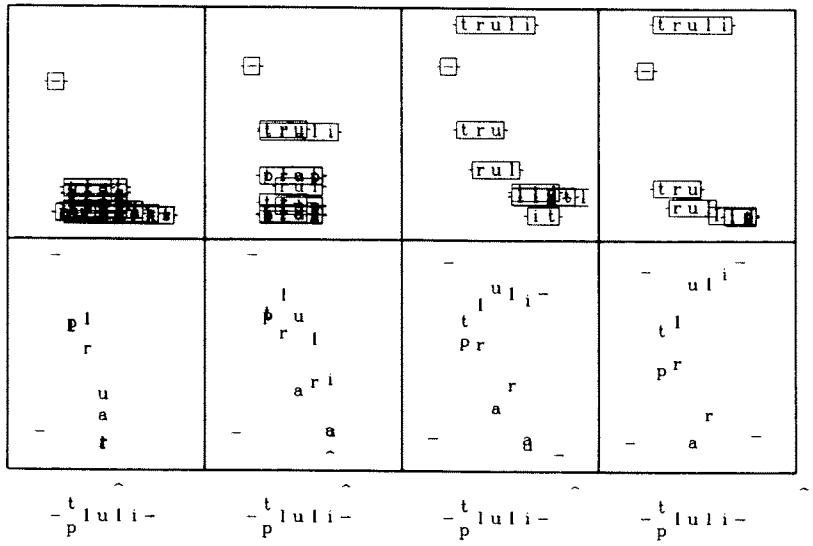

FIGURE 8. State of the Trace at several points in processing an ambiguous /p/-/t/ segment followed by /luli/.

Trading Relations and Categorical Perception

In the simulations considered thus far, phoneme identification is influenced by two different kinds of factors: featural and lexical. When one sort of information is lacking, the other can compensate for it. The image that emerges from these kinds of findings is of a system that exhibits great flexibility by being able to base identification decisions on different sources of information. It is, of course, well established that within the featural domain each phoneme is generally signaled by a number of different cues and that human subjects can trade these cues off against each other. The TRACE model exhibits this same flexibility, as we shall detail shortly.

But there is something of a paradox. While the perceptual mechanisms exhibit great flexibility in the cues that they rely on for phoneme identification, they also appear to be quite "categorical" in nature. That is, they produce much sharper boundaries between phonetic categories than we might expect based on their sensitivity to multiple cues; and they appear to treat acoustically distinct feature patterns as perceptually equivalent, as long as they are identified as instances of the same phoneme.

In this section, we illustrate that in TRACE, just as in human speech perception, flexibility in feature interpretation coexists with a strong tendency toward categorical perception.

For these simulations, the model was stripped down to the essential minimum necessary so that the basic mechanisms producing cue trade-offs and categorical perception could be brought to the fore. The word level was eliminated altogether, and at the phoneme level there were only three phonemes, /a/, /g/, and /k/, plus silence (/-/). From these four items, inputs and percepts of the form /-ga-/ and /-ka-/ could be constructed. The following additional constraints were imposed on the feature specifications of each of the phonemes: (a) the /a/ and /-/ had no featural overlap with either /g/ or /k/ so that neither /a/ nor /-/ would bias the activations of the /g/ and /k/ phoneme units where they overlapped with the consonant in time; (b) /g/ and /k/ were identical on five of the seven dimensions and differed only on the remaining two dimensions.

The two dimensions which differentiated /g/ and /k/ were voice onset time (VOT) and the onset frequency of the first formant (F1OF). These dimensions replaced the voicing and burst amplitude dimensions used in all of the other simulations. Figure 9 illustrates how F1OF tends to increase as voice onset time is delayed.

Trading relations. TRACE quite naturally tends to produce trading relations between features since it relies on the weighted sum of the

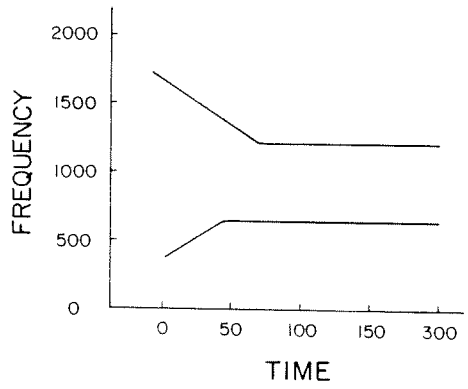

FIGURE 9. Schematic diagram of a syllable that will be heard as /ga/ or /ka/, depending on the point is the syllable at which voicing begins. Before the start of voicing, F2 (top curve) and F3 are energized by aperiodic noise sources, and F1 is "cut back" (the noise source has little or no energy in this range). Because of the fact that F1 rises over time after syllable onset (as the vocal tract moves from a shape consistent with the consonant into a shape consistent with the vowel), its frequency at the onset of voicing is higher for later values of VOT. Parameters used in constructing this schematic syllable are derived from Kewley-Port (1982).

excitatory inputs to determine how strongly the input will activate a particular phoneme unit. All else being equal, the phoneme unit 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. Summerfield and Haggard (1977) did this for VOT and F1OF and found the typical result, namely, that the value of one cue that gives rise to 50% choices of /k/ was affected by the value of the other cue: The higher the value of F1OF, the shorter the value of VOT needed for 50% choices of /k/. Unfortunately, they did not present full curves relating phoneme identification to the values used on each of the two dimensions. In lieu of this, we present curves in Figure 10 from a classic trading relations experiment by Denes (1955). Similar patterns of results have been reported in other studies, using other cues (e.g., Massaro, 1981), though the transitions are often somewhat steeper.

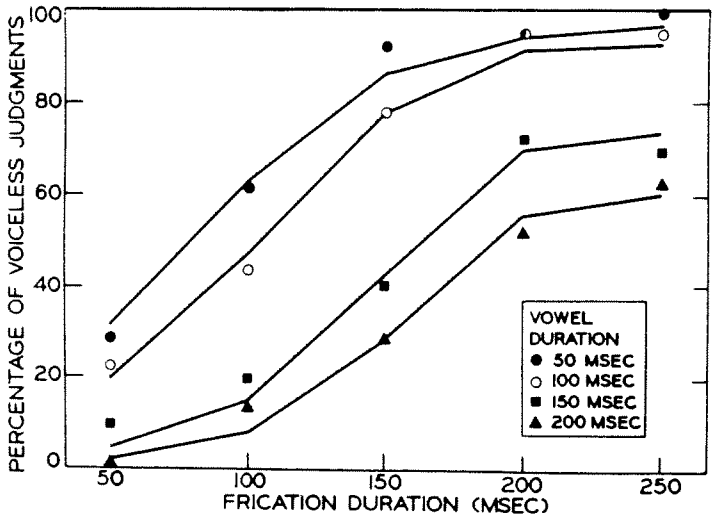

FIGURE 10. Results of an experiment demonstrating the trade-off between two cues to the identity of /s/ and /z/. Data from Denes, 1955, reprinted and fitted by the model of Massaro and Cohen. (From "The Contribution of Voice-Onset Time and Fundamental Frequency as Cues to the /zi/-/si/ Distinction" by D. W. Massaro and M. M. Cohen, 1977, *Perception & Psychophysics*, *22*, p. 374. Copyright 1977 by the Psychonomic Society. Reprinted by permission.)

To demonstrate that TRACE would simulate the basic tradeoff effect, we generated a set of 25 intermediate phonetic segments made up by pairing each of five different intermediate patterns on the VOT dimension with each of five different intermediate patterns on the F1OF dimension. The different feature patterns used on each dimension are shown in Figure 11, along with the canonical feature patterns for /g/ and /k/ on each of the two dimensions. On the remaining five dimensions, the intermediate segments all had the common canonical feature values for /g/ and /k/.

The model was tested with each of the 25 stimuli, preceded by silence (/-/) and followed by /a-/. 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. At the end of the 60th time slice, we recorded the activation of the units for /g/ and /k/ in Time-Slice 12 and the probability of choosing /g/ based on these activations. It makes no difference to the general pattern of the results if a different decision time is used.

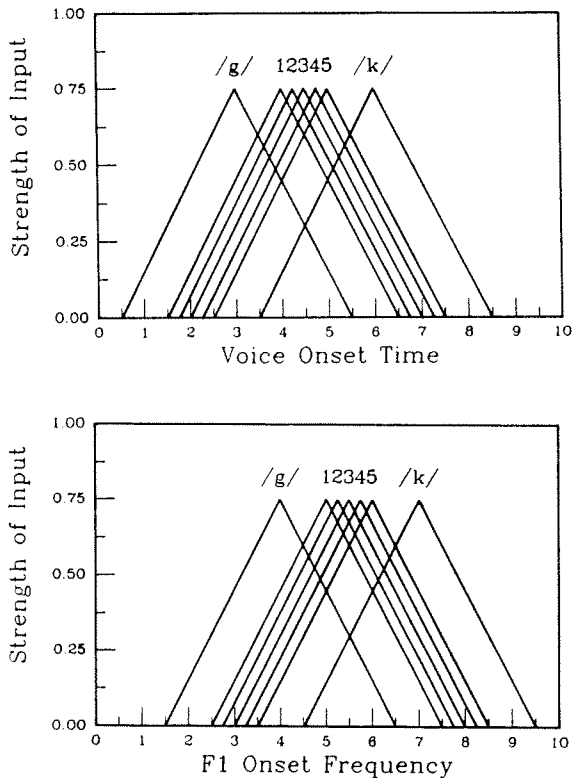

FIGURE 11. Canonical feature level input for /g/ and /k/, on the two dimensions that distinguish them, and the patterns used for the five intermediate values used in the trading relations simulation. Along the abscissa of each dimension, the nine units for the nine different value ranges of the dimension are arrayed. The curves labeled /g/ and /k/ indicate the relative strength of the excitatory input to each of these units produced by the indicated phoneme. The canonical curves also indicate the strengths of the feature to phoneme connections for /g/ and /k/ on these dimensions.

Response probabilities computed using the formulas given earlier are shown in Figure 12 for each of the 25 conditions of the experiment. The pattern of results is quite similar to that obtained in Denes' (1955) experiment on the /s/-/z/ continuum. The contribution of each cue is approximately linear and additive in the middle of the range, and the curves flatten out at the extremes, as in the Denes (1955) experiment. More importantly, the model's behavior exhibits the ability to trade one cue off against another. In terms of Summerfield and Haggard's measure, the value of VOT needed to achieve 50% probability of reporting

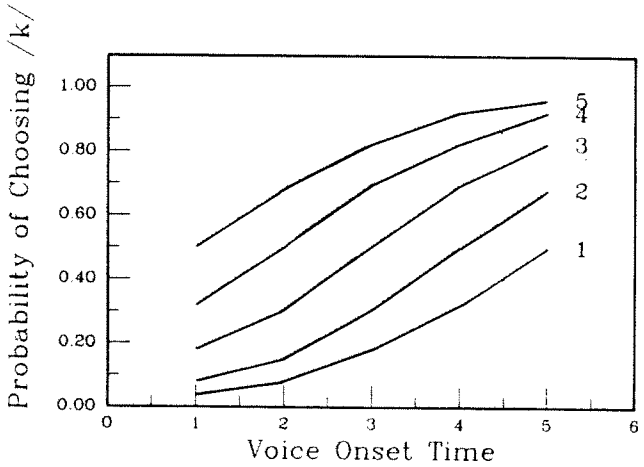

FIGURE 12. Simulated probability of choosing /k/ at Time-Slice 60, for each of the 25 stimuli used in the trading relations simulation experiment. Numbers next to each curve refer to the intermediate pattern on the FIOF continuum used in the 5 stimuli contributing to each curve. Higher numbers correspond to higher values of FIOF.

/k/, we can see that the VOT needed increases as the FIOF decreases, just as these investigators found.

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 /g/ to /k/ along the VOT and FIOF dimensions than we would expect from the slight changes in the relative excitation of the /g/ and /k/ units. 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 used in the simulations.

Eleven different consonant feature patterns were used, embedded in the same simulated /-_a-/ context as in the trading relations simulation. The stimuli varied from very low values of both VOT and FIOF, more extreme than the canonical /g/, through very high values on both dimensions, more extreme than the canonical /k/. All the stimuli were spaced equal distances apart on the VOT and FIOF dimensions. The locations of the peak activation values on each of these two continua are shown in Figure 13.

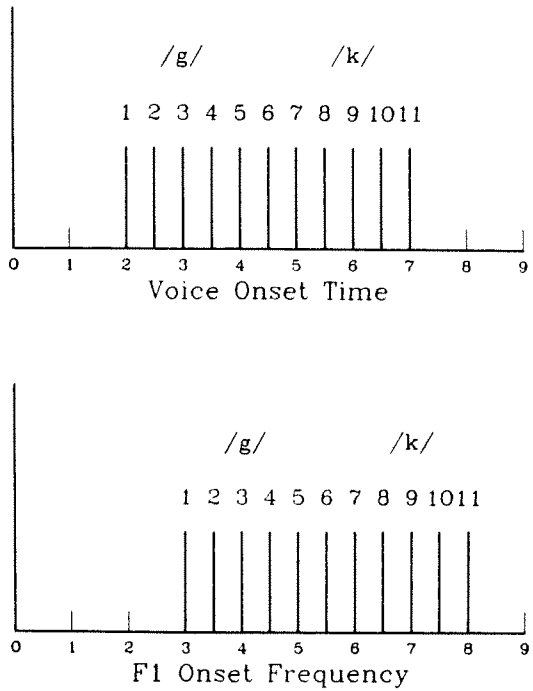

FIGURE 13. Locations of peak activations along the VOT and FIOF dimensions, for each of the 11 stimuli used in the categorical perception simulation.

Figure 14 indicates the relative initial bottom-up activation of the /g/ and /k/ phoneme units for each of the 11 stimuli used in the simulation. The first thing to note is that the relative bottom-up excitations of the two phoneme units differ only slightly. For example, the canonical feature pattern for /g/ sends 75% as much excitation to /g/ as it sends to /k/. The feature pattern two steps toward /g/ from /k/ (stimulus number 5), sends 88% as much activation to /g/ as to /k/.

The figure also indicates, in the second panel, the resulting activations of the units for /g/ and /k/ at the end of 60 cycles of processing. The slight differences in net input have been greatly amplified, and 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.

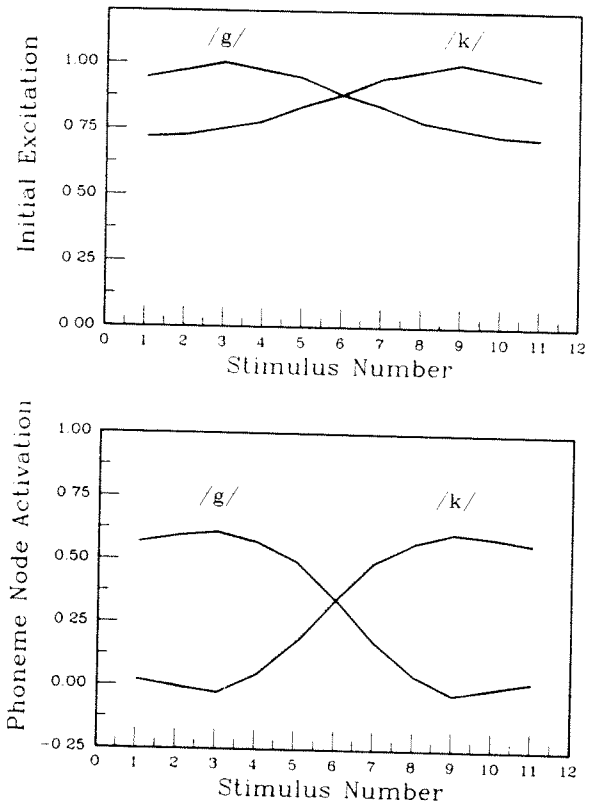

FIGURE 14. Effects of competition on phoneme activations. The first panel shows relative amounts of bottom-up excitatory input to /g/ and /k/ produced by each of the 11 stimuli used in the categorical perception simulation. The second panel shows the activations of units for /g/ and /k/ at Time-Cycle 60. Stimuli 3 and 9 correspond to the canonical /g/ and /k/, respectively.

Once one phoneme is slightly more active 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 has been noted many times (Grossberg, 1976; Levin, 1976; McClelland & Rumelhart, 1981). 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 14 are shown in Figure 15 along with the ABX discrimination function, which

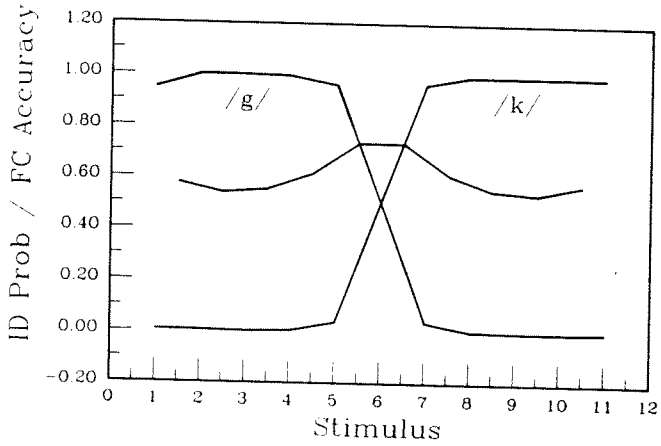

FIGURE 15. Simulated identification functions and forced-choice accuracy in the **ABX** task.

will be discussed later. The identification functions are even sharper than the activation curves; there is only a 4% chance that the model will choose /k/ instead of /g/ for Stimulus 5, for which /k/ receives 88% as much bottom-up support as /g/. 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 units, is a linear function of the difference in activation of the two units. When the activations are far enough apart, d' will be sufficient to ensure near-100% correct performance, even though both units have greater than zero activation. [2]

In TRACE, the categorical output of the model comes about only after an interactive competition process that greatly sharpens the differences in the activation of the detectors for the relevant units. This interactive process takes time. In the simulation results reported here, we assumed that subjects waited a fixed time before responding. But, if we assume that subjects are able to respond as soon as the response strength ratio reaches some critical level, we would find that subjects

[2] Many readers will note that the apparent sharpness of the identification functions shown in Figure 15 contrasts with the much shallower functions shown previously in the trading relations simulations. The reason for this is simply that the stimuli are spaced more closely together in the trading relations simulation than in the categorical perception case. This follows the standard experimental practice of emphasizing gradualness in trade-off experiments and sharpness in categorical perception experiments (Lane, 1965).

would be able to respond more quickly to stimuli near the prototype of each category than they can to stimuli near the boundary. This is exactly what was found by Pisoni and Tash (1974).

There is another aspect to categorical perception as exhibited by TRACE. This is the fact that 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. This allows the model to account for poor within-category discrimination and good between-category discrimination—the second hallmark of categorical perception.[3] The way it works is this. When a feature pattern comes in, it sends more excitation to some phoneme units than others; as they become active, they begin to compete, and one gradually comes to dominate the others. 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 units very briefly; the original pattern of activation produced by the phoneme units 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.

This effect is illustrated in Figure 16, which shows how different pairs of 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 /g/ and /k/ are considered in the difference measure.

[3] 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 that 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 that are always assigned to the same category. TRACE II produces this kind of approximate categorical perception.

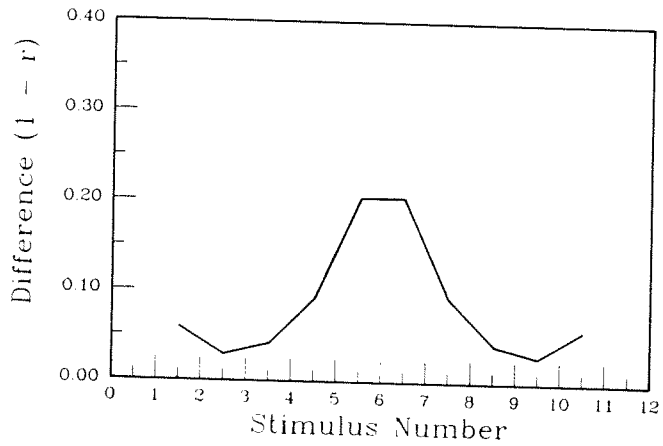

FIGURE 16. Differences between patterns of activation at the feature level at Cycle 60, for pairs of stimuli one step apart along the /g/-/k/ continuum used for producing the identification functions shown previously in Figure 15.

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 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. The resulting response probabilities are shown in Figure 15.

Basically, the figures show 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.

Like the completion process considered earlier, the process of "canonicalization" of the representation of a speech sound via the feedback mechanism takes time. During this time, two things are happening: One is that the activations initially produced by the speech input are decaying; another is that the feedback, which drives the representation toward the prototype, is building up. In the simulations, we allowed a considerable amount of time for these processes before computing similarities of different activation patterns to each other. Obviously, if we had left less time, there would not have been as much of an opportunity for these forces to operate. Thus, TRACE is in agreement with the finding that there tends to be an increase in within-category discrimination when a task is used that allows subjects to base their responses on judgments of the similarity of stimuli spaced closely together in time (Pisoni & Lazarus, 1974).

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 we would need to do is assume that the feature information that gives rise to phoneme identification is inaccessible, as proposed by the motor theory of speech perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), or is rapidly lost as proposed by the *dual code* model (Fujisaki & Kawashima, 1968; Massaro, 1975, 1981; Pisoni, 1973, 1975). The dual code model has had considerable success accounting for categorical perception data and accounts for all the aspects of categorical perception discussed thus far.

Both feedback models and dual code models can also accommodate the fact that vowels show less of a tendency toward categorical perception than consonants (Fry, Abramson, Eimas, & Liberman, 1962; Pisoni, 1973). It is simply necessary to assume that vowel features are more persistent than consonant features (Crowder, 1978, 1981; Fujisaki & Kawashima, 1968; Pisoni, 1973, 1975). However, the two classes of interpretations do differ in their predictions of performance in discriminating two stimuli, both away from the center of a category, but still within it. Here, TRACE tends to show greater discrimination than it shows between stimuli squarely in the middle of a category. Standard interpretations of categorical perception can account for increases in discriminability near the boundary between two categories (where identification may in fact be somewhat variable) by assuming that marginal stimuli are more likely to give rise to different category labels. But TRACE can account for increases in discriminability at extreme values of feature continua which would not give rise to different category labels. In TRACE, the reason for this increase in discriminability is that the activation of the appropriate item at the phoneme level is weaker, and therefore the feedback signal is weaker than it is when the input occurs near the center of the category. This results in less

canonicalization of the extreme stimuli and produces a W-shaped discrimination function, as shown in Figure 16. Few studies of categorical perception use stimuli that extend far enough into the extreme ranges of each phonetic category to observe reliable W-shaped curves; however, Samuel (1977) did carry out such a study and obtained just such W-shaped discrimination curves.

In summary, TRACE appears to provide a fairly accurate account of the phenomena of cue trade-offs and categorical perception of speech sounds. It accounts for categorical perception without relying on the notion that the phenomenon depends on read-out from an abstract level of processing; it assumes instead that the feature level, like other levels of the system, is subject to feedback from higher levels which actually changes the representation as it is being retained in memory, pushing it toward a canonical representation of the phoneme most strongly activated by the input.

Retuning of Phoneme Detectors by Context

In our simulations of trading relations, we have shown that the boundary between phonetic categories on one dimension can be affected by inputs on other dimensions. Other factors also influence the phoneme perceived as a result of particular featural input. The identity of phonemes surrounding a target phoneme, the rate of speech of a syllable in which a particular feature value occurs, as well as characteristics of the speaker and the language being spoken all influence the interpretations of features. See Repp and Liberman (1984) for a review of these effects.

In TRACE, we account for local, coarticulatory influences on phoneme identification by assuming that activations of phoneme units can modulate the feature to phoneme connections among units in adjacent time slices. This idea provides one way of implementing what Fowler (1984) has called a *factoring* of coarticulatory influences out of the pattern of activation produced by a feature pattern at the phoneme level. The modulation of connections can also account for the fact that phoneme boundaries shift as a function of local phonetic context (Mann & Repp, 1980). In simulations using TRACE I, we were able to improve the performance of the model in identifying the correct consonant at the beginning of a CV syllable from 79% correct without using connection modulation to 90% correct with connection modulation in place. Interestingly, the model is capable of generalizing from the connection strengths appropriate for the vowels it has been trained

on to other vowels, such as /e/, which fall between those on which it was trained. Figure 17 shows phoneme-level activations produced by the same token of the syllable /de/ with connection modulation turned on, in the upper panel, and off, in the lower panel. Though other units are activated in both cases, units for /d/ tend to dominate in the former case, but not the latter.

It has been suggested by J. L. Miller, Green, and Schermer (1984) that lexical effects and semantic and syntactic influences on phoneme identification may be due to a different mechanism than influences of such variables as speech rate and coarticulatory influences due to local phonetic context. The assumptions we have incorporated into TRACE

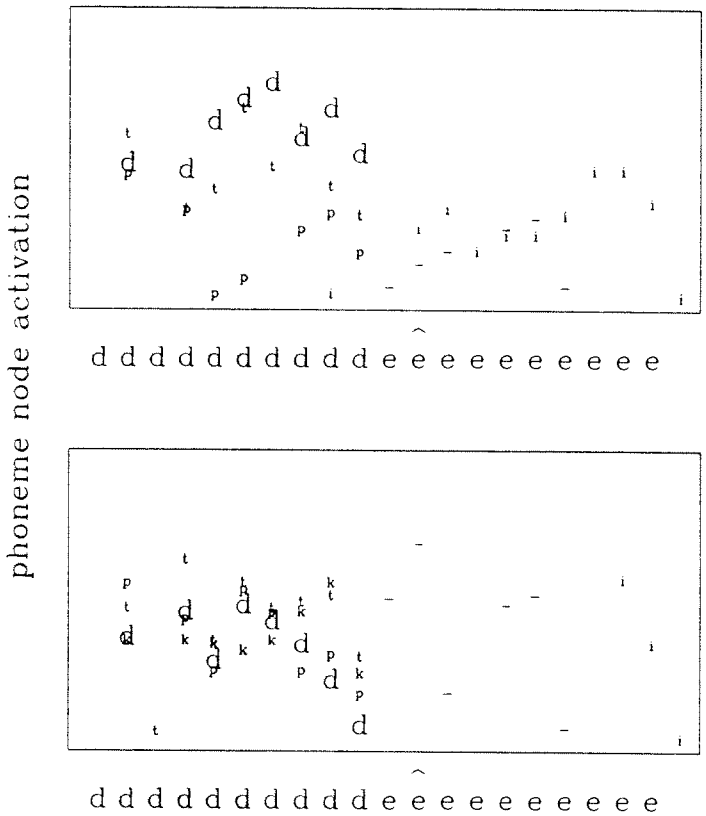

FIGURE 17. Activation of phoneme units resulting from the input /de/, with variable connection strengths enabled, in the upper panel, and disabled, in the lower panel. The /d/ units are depicted using a larger font just to increase their visibility.

make a similar distinction. Lexical effects are due to the additional source of input to the phoneme level provided by units at the word level. This is quite different from the connection modulation mechanism we have used to account for coarticulatory influences. In the discussion, we will consider ways of extending the connection modulation idea to accommodate effects of variations in rate and speaker parameters.

Summary of Phoneme Identification Simulations

We have considered a number of phenomena concerning the identification and perception of phonemes. These include lexical influences on phoneme identification and the lack thereof under certain circumstances; phonotactic rule effects on phoneme identification and the role of specific lexical items in influencing these effects; the integration of multiple cues to phoneme identity; and the categorical nature of the percept that results from this integration. We have also seen how connection modulation can be used to implement context-sensitive phoneme detectors, thereby allowing the model to improve its performance in identifying real speech and to account for effects of phonetic context on boundaries between phonemes. TRACE integrates all of these phenomena into a single account that incorporates aspects of the accounts offered for particular aspects of these results by other models. In the next section, we show how TRACE can also encompass a number of phenomena concerning the recognition of spoken words.

THE TIME COURSE OF SPOKEN WORD RECOGNITION

The study of spoken word recognition has a long history, and many models have been proposed. Morton's now-classic logogen model (Morton, 1969) was the first to provide an explicit account of the integration of contextual and sensory information in word recognition. Other models of this period (e.g., Broadbent, 1967) concentrated primarily on effects of word frequency. Until the midseventies, however, there was little explicit consideration of the time course of spoken word recognition. Several studies by Marslen-Wilson and his collaborators (Marslen-Wilson, 1973; Marslen-Wilson & Tyler, 1975), and by R. A. Cole and his collaborators (Cole, 1973; Cole & Jakimik, 1978, 1980) pioneered the investigation of this problem.

Marslen-Wilson's COHORT model (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978) of speech perception was based on this early work on the time course of spoken word recognition. The COHORT model was one of the sources of inspiration for TRACE for two main reasons. First, it provided an explicit account of the way top-down and bottom-up information could be combined to produce a word recognition mechanism that actually worked in real time. Second, it accounted for the findings of a number of important experiments demonstrating the *on-line* character of the speech recognition process. However, several deficiencies of the COHORT model have been pointed out, as we shall see.

Because TRACE was motivated in large part by a desire to keep what is good about COHORT and improve upon its weaknesses, we begin this section by considering the COHORT model in some detail. First we review the basic assumptions of the model, then consider its strengths and weaknesses.

There appear to be four basic assumptions of the COHORT model:

1. The model uses the first sound (in Marslen-Wilson & Tyler, 1980, the initial consonant-cluster-plus-vowel) of the word to determine which words will be in an initial cohort or candidate set.

2. Once the candidate set is established, the model eliminates words from the cohort immediately, as each successive phoneme arrives, if the new phoneme fails to match the next phoneme in the word. Words can also be eliminated on the basis of semantic constraints, although the initial cohort is assumed to be determined by acoustic input alone.

3. Word recognition occurs immediately, as soon as the cohort has been reduced to a single member; in an auditory lexical decision task, the decision that an item is a nonword can be made as soon as there are no remaining members in the cohort.

4. Word recognition can influence the identification of phonemes in a word only after the word has been recognized.

There is a considerable body of data that supports various predictions of the COHORT model. It has been observed in a variety of paradigms that lexical influences on phoneme identification responses are much greater later in words than at their beginnings (Bagley, 1900; R. A. Cole & Jakimik, 1978, 1980; Marslen-Wilson, 1980; Marslen-Wilson & Welsh, 1978). We considered some of this evidence in earlier sections. Another important finding supporting COHORT is the fact that the reaction time to decide that an item is a nonword is constant when measured from the occurrence of the first phoneme that rules out the last remaining word in the cohort (Marslen-Wilson, 1980).

Perhaps the most direct support for the basic word-recognition assumptions of COHORT comes from the gating paradigm, introduced first by Grosjean (1980). In this paradigm, subjects are required to guess the identity of a word after hearing successive presentations of the word. The first presentation is cut off so that the subject hears only the first N msec (N = 30 to 50 in different studies). Later presentations are successively lengthened in N msec increments until eventually the whole word is presented. The duration at which half the subjects correctly identify the word is called the *isolation point*. Considerably more input is required before subjects are reasonably sure of the identity of the word; that point is termed the *acceptance point*. Grosjean's initial study confirmed many basic predictions of COHORT, though it also raised a few difficulties for it (see below). In a more recent study using the same method, Tyler and Wessels (1983) carried out a very close analysis of the relation between the empirically determined isolation point and the point at which the input the subject has received is consistent with one and only one remaining item—the point at which recognition would be expected to occur in the COHORT model. They report that the isolation point falls very close to this theoretically derived recognition point, strongly supporting the basic immediacy assumptions of the COHORT model.

It should be noted that the gating task is not a timed task, and so it does not provide a direct measure of what the subject knows as the speech input is unfolding. However, it is now in fairly wide use, and Cotton and Grosjean (1984) have established that the basic patterns of results obtained in Grosjean's (1980) pioneering gating experiment do not depend on the presentation of successively longer and longer presentations of the same stimulus.

A dilemma for COHORT. Though the COHORT model accounts for a large body of data, there are several difficulties with it. We consider first the one that seems the most serious: as stated, COHORT requires accurate, undistorted information about the identity of the phonemes in a word up to the isolation point. Words cannot enter into consideration unless the initial consonant-cluster-plus-vowel is heard, and they are discarded from it as soon as a phoneme comes along that they fail to match. No explicit procedure is described for recovering words into the cohort once they have been excluded from it or when

the beginning of the word is not accurately perceived due to noise or elision.

These aspects of COHORT make it very difficult for the model to explain recognition of words with distorted beginnings, such as *dwibble* (Norris, 1982), or words whose beginnings have been replaced by noise (Salasoo & Pisoni, 1985). From a computational point of view, this makes the model an extremely brittle one; in particular, it fails to deal with the problem of noise and underspecification which is so crucial for recognition of real speech (Thompson, 1984).

The recognizability of distorted items like *dwibble* might be taken as suggesting that all we need to do is liberalize the criterion for entering and retaining words in the cohort. Thus, the cohort could be defined as the set of words consistent with what has been heard or mild deviations (e.g., one or two features) from what has been heard. This would allow mild distortions like replacing /r/ with /w/ not to disqualify a word from the cohort. It would also allow the model to cope with cases where the beginning of the word is underspecified; in these cases, the initial cohort would simply be larger than in the case where the input clearly specified the initial phonemes.

However, there is still a problem. Sometimes we need to be able to rule out items that mismatch the input on one or two dimensions 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 lose 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 select the correct word as soon as an undistorted input specifies it uniquely, to account for the Tyler and Wessels results. On the other hand, we do not want the model to completely eliminate possibilities that might later turn out to be correct. We shall shortly see that TRACE provides a way out of this dilemma.

Another problem for COHORT. Grosjean (1985) has recently pointed out another problem for COHORT, namely, the possibility that the subject may be uncertain about the location of the beginning of each successive word. A tacit assumption of the model is that the subject goes into the beginning of each word knowing that it is the beginning. In the related model of R. A. Cole and Jakimik (1980), this assumption is made explicit. Unfortunately, it is not always possible to know in advance where one word starts and the next word ends. As we discussed in the introduction, acoustic cues to juncture are not always reliable, and in the absence of acoustic cues, even an optimally efficient

mechanism cannot always know that it has heard the end of one word until it hears enough of the next to rule out the possible continuations of the first word.

What is needed, then, is a model that can account for COHORT's successes and overcome these two important deficiencies. The next two sections show that TRACE does quite well on both counts. The first of these sections examines TRACE's behavior in processing words whose beginnings and endings are clearly delineated for it by the presence of silence. The second considers the processing of multiword inputs, which the model must parse for itself.

One Word at a Time

In this section, we see how TRACE resolves the dilemma facing COHORT, in that it is immediately sensitive to new information but is still able to cope with underspecified or distorted word beginnings. We also consider how the model accounts for the preference for short-word responses early in processing a long word. The section concludes with a discussion of ways the model could be extended to account for word frequency and contextual influences.

Competition vs. bottom-up inhibition. TRACE deals with COHORT's dilemma by using competition rather than phoneme-to-word inhibition. The essence of the idea is simply this: Phoneme units have excitatory connections to all the word units they are consistent with. Thus, whenever a phoneme becomes active in a particular slice of the Trace, it sends excitation to all the word units consistent with that phoneme in that slice. The word units then compete with each other; items that contain each successive phoneme in a sequence dominate all others, but, if no word matches perfectly, a word that provides a close fit to the phoneme sequence can eventually win out over words that provide less adequate matches.

Consider, from this point of view, our two items *pleasant* and *blacelet* again. In the first instance, *pleasant* will receive more bottom-up excitation than *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 this 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, *bracelet* will eventually come to dominate the other possibilities.

This behavior of the model is illustrated using examples from its restricted lexicon in Figure 18. In one case, the input is *legal*, and the word *regal* is completely dominated by *legal*. In the other case, the input is *lugged*, and the word *rugged* eventually dominates because there is no word *lugged* (pronounced to rhyme with *rugged*—the word *lug* is not in the model's lexicon). Here, *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, but it does win out in the end.

It should be noted that the details of what word will be most strongly activated in such cases depend on a number of factors, including, in particular, the distinctiveness of mismatching phonemes. Also, it is possible to find cases in which a word that 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 or losing, but rather the fact that both tend to stay in the game. 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 in cases where the beginning of a word has been distorted. In earlier sections on phoneme processing, some of the simulations illustrate that the model is capable of recognizing words with underspecified (i.e., ambiguous) initial phonemes. In this section, we examine how well TRACE emulates the COHORT model in cases where the input is an undistorted representation of some particular word. In particular, we wanted to see how close TRACE would come to behaving in accord with COHORT's assumption that incorrect words are dropped from the cohort of active candidates as soon as the input diverges from them.

To examine this process, we considered the processing of the word *product* (/prad^ct/). Figure 19 shows the state of the Trace at various points in processing this word, and Figure 20 shows the response strengths of several units 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 the latter figure, the response strength of *product* is simply given as 1.0 at each time slice and the response strengths of units for other words are given relative to the strength of *product*. The curves shown are for the words *trot*, *possible*, *priest*, *progress*, and *produce*; these words differ from the word *product* (according

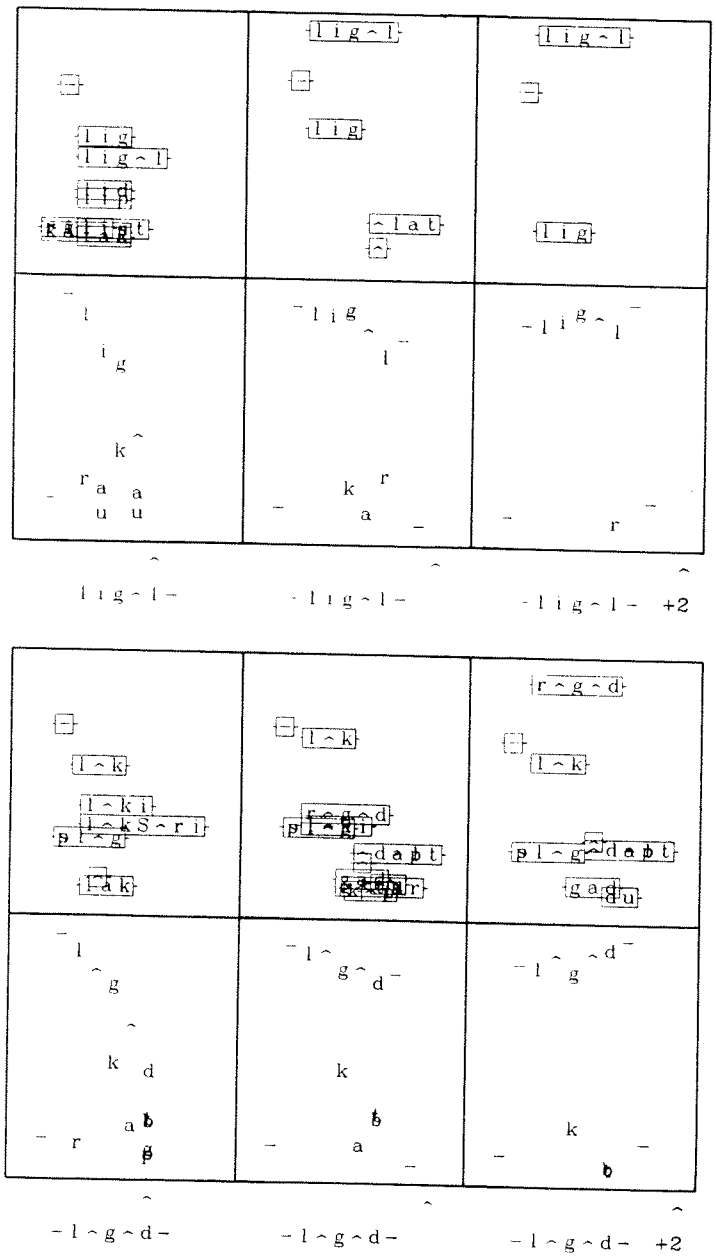

FIGURE 18. State of the Trace at three points during the processing of *legal* and *lugged*.

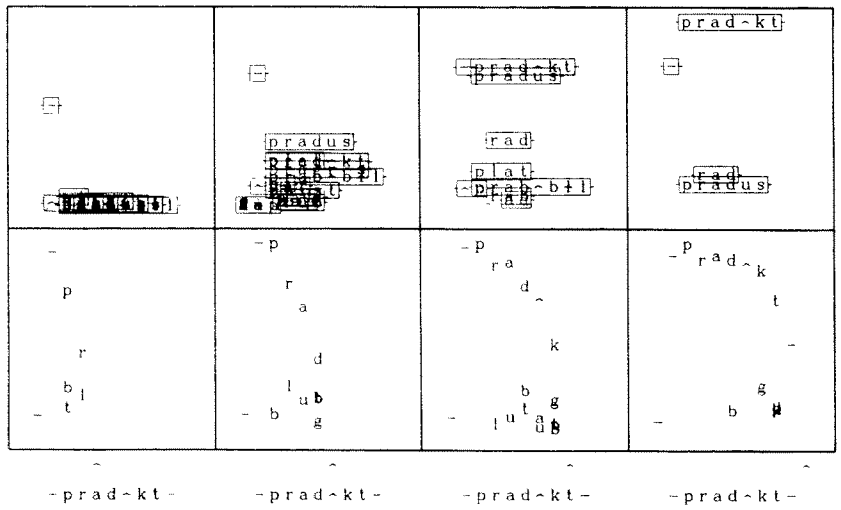

FIGURE 19. State of the Trace at various points in processing the word *product* (/prad^kt/).

to the simulation program's stressless encoding of them!) in the first, second, third, fourth, and fifth phonemes, respectively. Figure 20 shows that these items begin to drop out just after each successive phoneme comes in. Of course, there is nothing hard and fast or absolute about dropping a candidate in TRACE. What we see instead is that mismatching candidates simply begin to fade as the input diverges from them in favor of some other candidate. This is just the kind of behavior the COHORT model would produce in this case, though, of course, the drop-off would be assumed to be an abrupt, discrete event.[4]

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 (in phonemes, *produce* is one shorter than *product*).

[4] The data reported by Tyler and Wessels actually appears to indicate an even more immediate drop-off than is seen in this simulation. However, it should be remembered that the curves shown in Figure 20 are on-line response strength curves, and thus reflect the lags inherent in the percolation of input from the feature to the word level. The gating task, on the other hand, does not require subjects to respond on-line. If the input is simply turned off at the peak of each phoneme's input specification and then allowed to run free for a few cycles, the dropout point shifts even earlier.

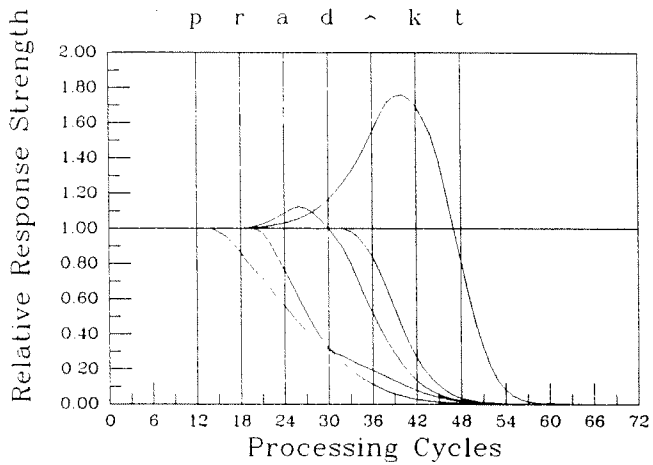

FIGURE 20. Response strengths of the units for several words relative to the response strength of the unit for *product* (/prad^kt/), as a function of time relative to the peak of the first phoneme that fails to match the word. The successive curves coming off of the horizontal line representing the normalized response strength of *product* are for the words *trot*, *possible*, *priest*, *progress*, and *produce*, respectively. In our lexicon they are rendered as /trat/, /pas^b^l/, /prist/, /pragr^s/, and /pradus/, respectively.

This advantage for shorter words is due to the competition mechanism. Recall that word units compete with each other in proportion to the overlap of the sets of time slices spanned by each of the words. Overlap is, of course, symmetrical, so long and short words inhibit each other to an equal extent; but longer words suffer more inhibition from other long words than short words do. For example, *progress* and *probable* inhibit *product* more than they inhibit *priest* and *produce*. Thus, units for longer words are generally subjected to extra inhibition, particularly early on in processing when many candidates are active, and so they tend to suffer in comparison to short words as a result.

We were at first somewhat disturbed by this aspect of the model's behavior, but it turns out to correspond quite closely with results obtained in experiments by Grosjean (1980) and Cotton and Grosjean (1984) using the gating paradigm. Both papers found that subjects hearing the beginnings of words like *captain* tended to report shorter words consistent with what they had heard (e.g., *cap*). However, we should observe that in the gating paradigm, when the word *captain* is truncated just after the /p/, it will sound quite a bit like *cap* followed by silence. In TRACE, this silence would activate silence units at the phoneme and word levels, and the word-level silence units would compete with units for words that extend into the silence. The silence will reinforce the preference of the model for short-word interpretations

because the detection of the silence will inhibit the detector for the longer word. Thus, there are actually two reasons why TRACE might favor short-word interpretations over long-word interpretations in a gating experiment. Whether human subjects show a residual preference for shorter interpretations over longer ones in the absence of a following silence during the course of processing is not yet clear from available data.

We should point out that the experimental literature indicates that the advantage of shorter words over longer ones holds only under the special circumstances of gated presentation and then only with early gates, when shorter words are relatively more complete than longer ones would be. It has been known for a long time that longer words are generally more readily recognized than shorter ones when the whole word is presented for identification against a background of noise (Licklider & Miller, 1951). Presumably, the reason for this is simply that longer words generally provide a larger number of cues than shorter words do, and hence are simply less confusable.

Frequency and context effects. There are, of course, other factors that influence when word recognition will occur, beyond those we have considered thus far. Two very important ones are word frequency and contextual predictability. The literature on these two factors goes back to the turn of the century (Bagley, 1900). Morton's (1969) logogen model effectively deals with several important aspects of this huge literature, though not with the time course of these effects.

We have not yet included either word frequency or higher-level contextual influences in TRACE, though, of course, we believe they are important. Word frequency effects could be accommodated, as they were in the interactive activation model of word recognition, in terms of variation in the resting activation level of word units or in terms of variation in the strength of phoneme-to-word connections. Contextual influences can be thought of as supplying activation to word units from even higher levels of processing than the word level. In this way, basic aspects of these two kinds of influences can be captured. We leave it to future research, however, to determine to what extent these elaborations of TRACE would provide a detailed account of the data on the roles of these factors. For now, we turn to the problem of determining where one word ends and the next one begins.

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, as we noted in the introduction. To recap our earlier argument, there are some cues in the speech stream, but as several investigators have pointed out (R. A. Cole & Jakimik, 1980; Grosjean & Gee, 1984; Thompson, 1984), they are not always sufficient, particularly in fluent speech. It would appear that there is an important role for lexical knowledge to play in determining where one word ends and the next word begins, as well as in identifying the objects that result from the process of segmentation. Indeed, as Reddy (1976) has suggested, segmentation and identification may simply be joint results of the mechanisms of word recognition.

R. A. Cole and Jakimik (1980) discuss these points and present evidence that semantic and syntactic context can guide segmentation in cases where the lexicon is consistent with two readings (*car go* vs. *cargo*). 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 out a sequence of words in continuous streams of phonemes. Word identification and segmentation emerge together from the interactive activation process as part and parcel of the process of word activation.

This section considers several aspects of the way in which word segmentation emerges from 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 units and about competitive inhibition between word units. First, the extent to which a particular phoneme excites a particular word unit is independent of the length of the word. Second, the extent to which a particular word unit will inhibit another word unit is proportional to the temporal overlap of the two word units. This means that words which do not overlap in time will not inhibit each other but will gang up on other words that 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 utterances *she can't* and *secant*. 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 seems to be two words and the second seems like only one.

To see what TRACE II would do with single and multiple word inputs, we ran simulation experiments with each individual word in the main 211-word lexicon preceded 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. We stress, of course, that real speech would tend to contain cues that would mark word boundaries in many cases; the experiment is simply designed to show what TRACE can do in cases where these cues are lacking.

With the individual words, TRACE made no mistakes—that is, by a few slices after the 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 21. The stream /parti/ might be either one word (*party*) or two (*par tea*, or *par tee*—the model knows of only one word pronounced /ti/). At early points in processing the word, *par* dominates over *party* and other longer words for reasons discussed in the previous section. By the time the model has had a chance to process the end of the word, however, *party* comes to dominate.

Why does a single longer word eventually win out over two shorter ones in TRACE? There are two main reasons. First of all, a longer word eventually receives more bottom-up support than either shorter word, simply because there are more phonemes activating the longer word than the shorter word. The second reason has to do with the

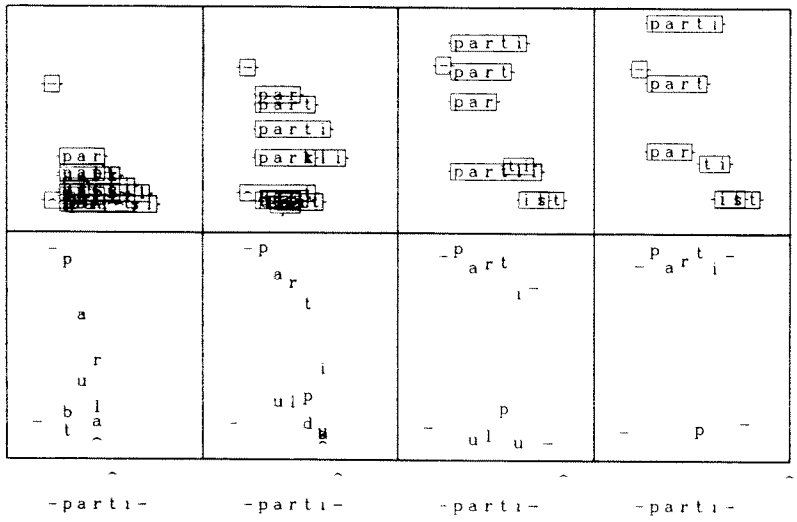

FIGURE 21. The state of the Trace at various points during processing of /parti/.

sequential nature of the input. In the case of /parti/, by the time the /ti/ is coming in, the word *party* is well enough established that it keeps /ti/ from getting as strongly activated as it would otherwise. 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 could be tested using the gating paradigm or a cross-modal priming paradigm such as the one used by Swinney (1982).

However, it should be noted that this aspect of the behavior of the model can be overridden if there is bottom-up information favoring the two-word interpretation. Currently, this can only happen in TRACE through the insertion of a brief silence between the *par* and the *tea*. As shown in Figure 22, this results in *par* and *tea* dominating all other word candidates.

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 tea*, as shown in Figure 22. Note that other words, such as *art*, that span a portion of the input, are less successful than either *bar* or *tea*. The reason is that the interpretations *bar* and *art* overlap with each other, and *art* and *tea* overlap with each other, but *bar* and *tea* do not overlap. Thus, *art* receives inhibition from both *bar*

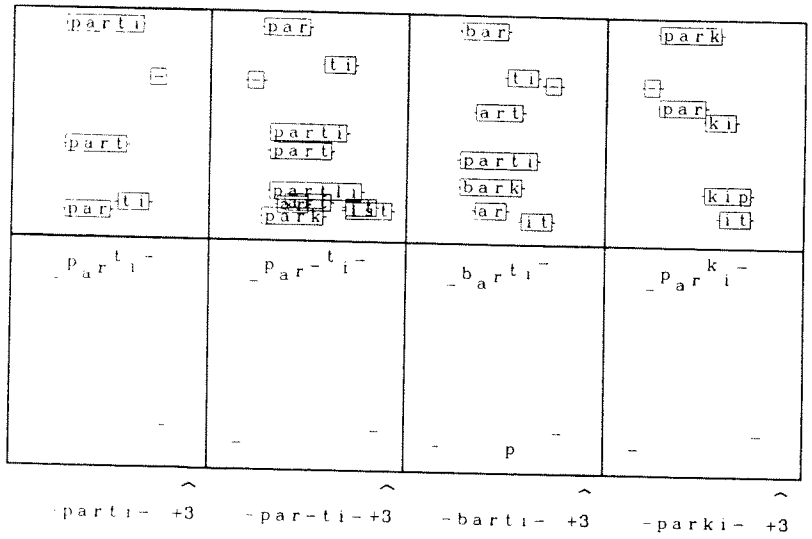

FIGURE 22. State of the Trace after processing the streams /parti/, /par-ti/, /barti/, and /parki/.

and *tea*, while *bar* and *tea* each receive inhibition only from *art*. Thus, two words that do not overlap with each other can gang up on a third word that each overlaps with partly and drive it out.

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 189 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* for *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 nonoverlapping 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 22.

This aspect of TRACE II's behavior indicates that the present version of the model is far from the final word on word segmentation. 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 in all cases without the aid of other cues.

Some readers may be troubled by a mechanism that does not insist upon a parse in which each phoneme is covered by one and only one word. Actually, though, this characteristic of the model 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 speakers tend to signal the doubling in careful speech, the cues to single vs. double consonants are not always sufficient for disambiguation, as is clear when strings with multiple interpretations are used as stimuli. For example, an utterance intended as *no notion* will sometimes be heard as *known notion* (Nakatani & Dukes, 1977). The model is not inclined to suppress activations of partially overlapping words, even when a nonoverlapping parse is available. This behavior of TRACE is illustrated with /bˆstap/ (*bus top* or *bus stop*) in Figure 23. In this case, higher levels could provide an additional source of information that would help the model choose between overlapping and nonoverlapping interpretations.

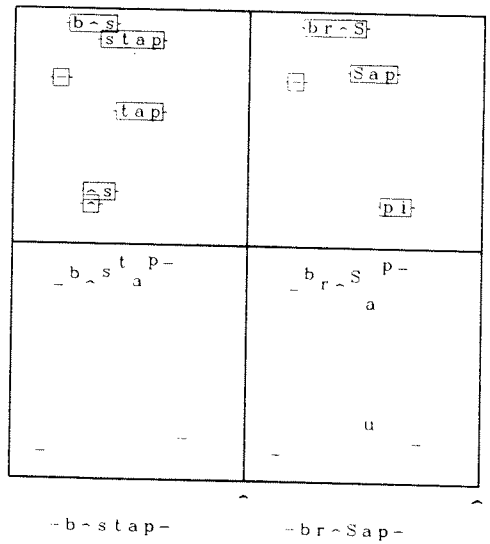

FIGURE 23. State of the Trace at the end of the streams /bustap/ (*bus stop* or *bus top*) and /bruSap/ (*brush shop*).

Thus far in this section, we have considered the general properties of the way in which TRACE uses lexical information to segment a speech stream into words, but we have not considered much in the way of empirical data that these aspects of the model shed light on. However, there are two findings in the literature which can be interpreted in accordance with TRACE's handling of multiword speech streams.

Where does a nonword end? A number of investigators (e.g., R. A. Cole & Jakimik, 1980) have suggested that when one word is identified, its identity can be used to determine where it ends and, therefore, where the next word begins. In TRACE, the interactive activation process can often establish where a word will end even before it actually does end, particularly in the case of longer words or when 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 target phonemes at the beginning of a word preceded by a nonword than at the beginning of a word

preceded by a word. For example, responses to detect word-initial /d/ were faster in stimuli such as the following:

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

than they were when the word preceding the target phoneme (in this case, *government*) was replaced by a nonword such as *gatabont*. It should be noted that the targets were specified as word-initial segments. Therefore, the subjects had not only to identify the target phoneme; they had to determine that it fell at the beginning of a word as well. The fact that reaction times were faster when the target was preceded by a word suggests that subjects were able to use their knowledge of where the word *government* ends to help them determine where the next word begins.

An example of how TRACE allows one word to help establish where its successor begins is illustrated in Figure 24. 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 initial /t/ in *target* becomes active at the phoneme level (second panel in the upper part of the figure), so there is no ambiguity about the fact that this /t/ is at the beginning of the next word. (The decision mechanism would, of course, be required to note that the model had established the location of the end of the preceding word. We have not yet incorporated explicit assumptions about how this would be done.) In the second case, words beginning and ending at a number of different places, including some that overlap with the location of the /t/, are partly activated. Thus, a listener would have to wait until the input 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 many cases there will be cues in the manner of pronunciation and the syllabification of the input 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 helping listeners determine where the beginning of the next word must be.

The long and short of word identification. One problematic 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*, though *tar* will tend to be

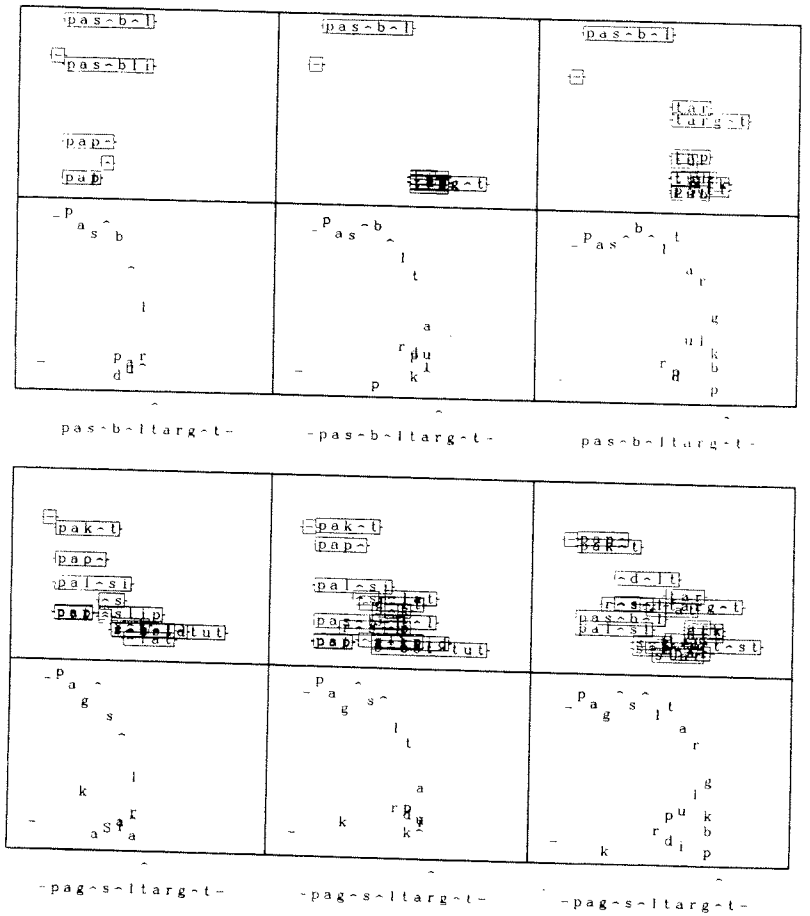

FIGURE 24. State of the Trace at several points during the processing of *possible target* and *pagusle target*.

the preferred hypothesis at this point, we do not have enough information to say unequivocally that the *tar* will not turn out to be *target* or *tarnished* or one of several other possibilities. It is only after 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/ (*tar box*), is illustrated in Figure 25. Though *tar* is somewhat more active than the longer word *target* when the /r/ is coming in, it is only when the word *box* emerges as the interpretation of the phonemes following *tar* that the rival *target* finally fades as a serious contender.

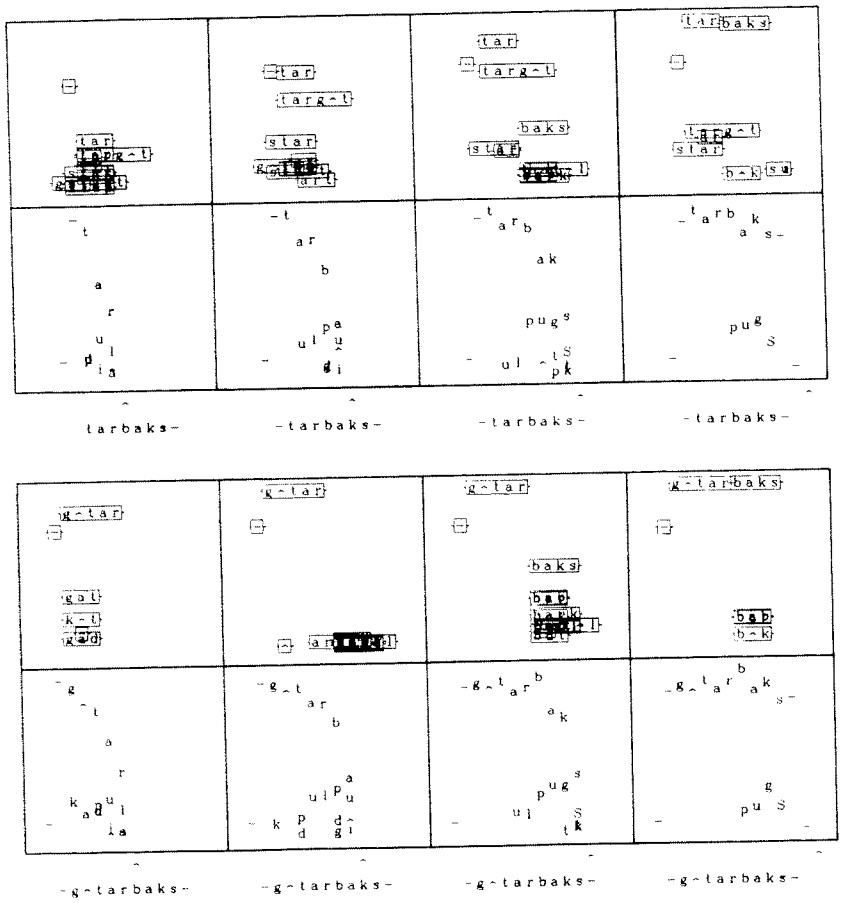

FIGURE 25. State of the Trace at several points in processing *tar box* and *guitar box*.

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 well before the end of the word. An illustration of this situation is provided by a simulation using the utterance *guitar box* /g^tarbaks/. By the time the /r/ has registered, *guitar* is clearly dominant at the word level and can be unambiguously identified without further ado.

Recently, an experiment by Grosjean (1985) 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. With longer words, subjects could usually guess the word correctly well before the end of the word; and by the end of the word, they were quite sure of the word's identity. With monosyllabic words, on the other hand, many of the words could not be identified correctly until well into the next word. On average, subjects were not sure of the word's identity until about the end of the next word or the beginning of the one after. As Grosjean (1985) points out, a major reason for this is simply that the spoken input often does not uniquely specify the identity of a short word. In such cases, the perceptual system is often forced to process the short word and its successor at the same time.

Recognizing the words in a short sentence. One last example of TRACE II's performance in segmenting words is illustrated in Figure 26. The figure shows the state of the Trace at several points during the processing of the stream /SiS^t^baks/. By the end, the words of the phrase *she shut a box*, which fits the input perfectly with no overlap, dominate all others.

This example illustrates how far it is sometimes possible to go in parsing a stream of phonemes into words without even considering syntactic and semantic constraints, or stress, syllabification, and juncture cues to word identification. The example also illustrates the difficulty the model has in perceiving short, unstressed words like *a*. This is, of course, just an extreme version of the difficulty the model has in processing monosyllabic words like *tar* and is consistent with Grosjean's data on the difficulty subjects have with identifying short words. In fact, Grosjean and Gee (1984) report pilot data indicating that these difficulties are even more severe with function words like *a* and *of*. It should be noted that TRACE makes no special distinction between content and function words per se, and neither do Grosjean and Gee. However, function words are usually unstressed and considerably shorter than content words. Thus, it is not necessary to point to any special mechanisms for closed versus open class morphemes to account for Grosjean and Gee's results.

Summary of Word Identification Simulations

While phoneme identification has been studied for many years, data from on-line studies of word recognition is just beginning to accumulate. There is an older literature on accuracy of word identification in noise, but it has only been quite recently that useful techniques have been developed for studying word recognition in real time.

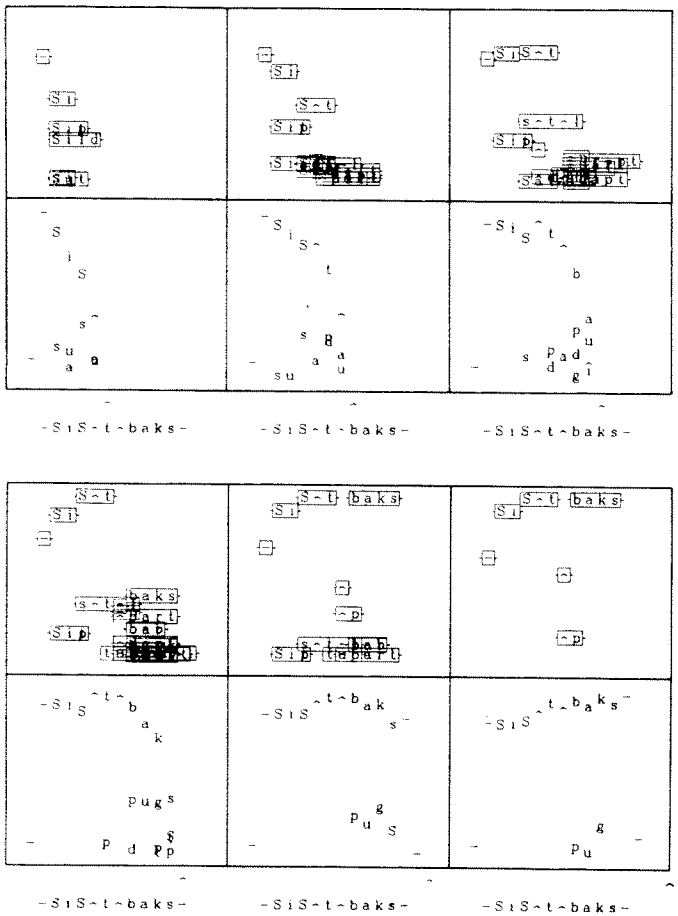

FIGURE 26. The state of the Trace at several points during the processing of the stream /SiS^t^baks/ (*she shut a box*).

What evidence there is, though, indicates the complexity of the word identification process. While the word identification mechanism is sensitive to each new incoming phoneme as it arrives, it is nevertheless robust enough to recover from underspecification or distortion of word beginnings. And, it appears to be capable of some simultaneous processing of successive words in the input stream. TRACE appears to capture these aspects of the time course of word recognition. In these respects, it improves upon the COHORT model, the only previously extant model that provides an explicit account of the on-line process of

word recognition. And, the mechanisms it uses to accomplish this are the same ones that it used for the simulations of the process of phoneme identification described in the preceding section.

GENERAL DISCUSSION

Summary of TRACE's Successes

In this chapter, we have seen that TRACE can account for a number of different aspects of human speech perception. We begin by listing the major correspondences between TRACE and what we know about the human speech understanding process.

1. TRACE, like humans, uses information from overlapping portions of the speech wave to identify successive phonemes.

2. The model shows a tendency toward categorical perception of phonemes, as do human subjects. The model's tendency toward categorical perception is affected by many of the same parameters that affect the degree of categorical perception shown by human subjects; in particular, the extent to which perception will be categorical increases with time between stimuli that must be compared.

3. The model combines feature information from a number of different dimensions and exhibits cue trade-offs in phoneme identification.

4. The model augments information from the speech stream with feedback from the lexical level in reaching decisions about the identity of phonemes. These lexical influences on phoneme identification occur in conditions similar to those in which lexical effects have been reported, but do not occur in conditions in which these effects have not been obtained.

5. Like human subjects, the model exhibits apparent phonotactic rule effects on phoneme identification, though it has no explicit representation of the phonotactic rules. The tendency to prefer phonotactically regular interpretations of ambiguous phonemes can be overridden by particular lexical items, just as it can in the human perceiver.

6. Our simulations with TRACE I show that the model is able to use activations of phoneme units in one part of the Trace to adjust the connection strengths determining which features will activate which phonemes in adjacent parts of the Trace. In this way the model can adjust as human subjects do to coarticulatory influences on the acoustic properties of phonemes (Fowler, 1984; Mann & Repp, 1980).

7. In processing unambiguous phoneme sequences preceded by silence, the model exhibits immediate sensitivity to information favoring one word interpretation over another. It shows an initial preference for shorter words relative to longer words, but eventually a sequence of phonemes that matches a long word perfectly will be identified as that word, overturning the initial preference for the short-word interpretation. These aspects of the model are consistent with human data from gating experiments.

8. Though the model is heavily influenced by word beginnings, it can recover from underspecification or distortion of a word's beginning.

9. The model can use its knowledge of the lexicon to parse sequences of phonemes into words and to establish where one word ends and the next one begins when cues to word boundaries are lacking.

10. Like human subjects, the model sometimes cannot identify a word until it has heard part of the next word. Also like human subjects, it can better determine where a word will begin when it is preceded by a word rather than a nonword.

11. The model does not demand a parse of a phoneme sequence that includes each phoneme in one and only one word. This allows it to cope gracefully with elision of phonemes at word boundaries. It will often permit several alternative parses to remain available for higher-level influences to choose among.

There is, of course, more data on some of these points than others. It will be very interesting to see how well TRACE will hold up against the data as further empirical studies are carried out.

Reasons for TRACE's Successes

We think there are two main reasons why TRACE has worked so well. The first is its use of massively parallel, interactive processing. The second is the Trace architecture. We do not believe that the model would have worked without both of these characteristics. The Trace provides a processing structure that lays out the hypotheses about the contents of an utterance in a way that captures directly the task to be performed—to find an interpretation of the utterance consisting of a sequence of units on each of several processing levels. Appropriate competition within levels and mutual facilitation between levels is quite naturally arranged in such a situation. This is important, but it would not work without the parallel, interactive processing that is provided by the PDP framework.

There is evidence to support both parts of this point. One source comes from our early attempts to model speech perception without adopting the Trace architecture. Our early model (described in Elman & McClelland, 1984) failed to provide a straightforward representation of the temporal structure of the speech stream.

It is a commonplace observation in the field of artificial intelligence that success in modeling depends on having the right representation, and Marr (1982) made very much of this point. But it is also true that one must have the right kind of processing system to exploit the architecture. The fact that the HEARSAY speech understanding system was not terribly successful attests to this. HEARSAY had the right architecture—a better one, in fact, in some ways, than we have in the current version of TRACE (see Chapter 16). But HEARSAY lacked the massively parallel, interactive capabilities of PDP models.

Some Deficiencies of TRACE

Although TRACE has had a number of important successes, it also has a number of equally important deficiencies. One fundamental deficiency is the fact that the model requires reduplication of units and connections, copying over and over again the connection patterns that determine which features activate which phonemes and which phonemes activate which words. One reason why this is a problem has to do with learning. Learning in PDP models involves tuning connections between pairs of units based on both of their states. This kind of learning is inherently *local* to the specific connections between the specific units involved. Given TRACE's 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 and 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 & 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.

All these deficiencies reflect the fact that the Trace consists of a large set of independent tokens of each feature, phoneme, and word unit. What appears to be called for instead 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 units, 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 as we argued just above. Thus, it will not do to simply abandon the Trace in favor of a single set of units 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. The next chapter describes a model of reading that has some of these characteristics. It uses connection information stored in a central PDP network to set connections in a processing structure much like the Trace, thereby effectively programming this structure in the course of processing. The next step in the development of TRACE is to apply these ideas to speech perception. Some comments about how this might be done are included at the end of the next chapter.

CONCLUSION

Our aim in this chapter has been to show that parallel distributed processing mechanisms provide a natural framework for developing models capable of meeting the computational challenges posed by

speech and of accounting for the data on human speech perception. The TRACE model does quite well on both counts. Though the architecture of the model is partially responsible for its successes, we have argued that the successes of the model depend at least as much on the parallel distributed processing operations that take place within the architecture. TRACE does have some limitations, but it successes so far have been quite encouraging. Just how easy it will be to overcome the limitations is a matter for future research.

ACKNOWLEDGMENTS

The work reported here was supported in part by a contract from the Office of Naval Research (N-00014-82-C-0374, NR 667-483), in part by a grant from the National Science Foundation (BNS-79-24062), and in part by a Research Scientist Career Development Award to the first author from the National Institute of Mental Health (5-K01-MH00385).