COGNITIVE PSYCHOLOGY 18, 1–86 (1986)

The TRACE Model of Speech Perception

JAMES L. MCCLELLAND

Carnegie–Mellon University

AND

JEFFREY L. ELMAN

University of California, San Diego

We describe a model called the TRACE model of speech perception. The model is based on the principles of interactive activation. Information processing takes place through the excitatory and inhibitory interactions of a large number of simple processing units, each working continuously to update its own activation on the basis of the activations of other units to which it is connected. The model is called the TRACE model because the network of units forms a dynamic processing structure called "the Trace," which serves at once as the perceptual processing mechanism and as the system's working memory. The model is instantiated in two simulation programs. TRACE I, described in detail elsewhere, deals with short segments of real speech, and suggests a mechanism for coping with the fact that the cues to the identity of phonemes vary as a function of context. TRACE II, the focus of this article, simulates a large number of empirical findings on the perception of phonemes and words and on the interactions of phoneme and word perception. At the phoneme level, TRACE II simulates the influence of lexical information on the identification of phonemes and accounts for the fact that lexical effects are found under certain conditions but not others. The model also shows how knowledge of phonological constraints can be embodied in particular lexical items but can still be used to influence processing of novel, nonword utterances. The model also exhibits categorical perception and

The work reported here was supported in part by a contract from the Office of Naval Research (N-00014-82-C-0374), in part by a grant from the National Science Foundation (BNS-79-24062), and in part by a Research Scientists Career Development Award to the first author from the National Institute of Mental Health (5-K01-MH00385). We thank Dr. Joanne Miller for a very useful discussion which inspired us to write this article in its present form. David Pisoni was extremely helpful in making us deal more fully with several important issues, and in alerting us to a large number of useful papers in the literature. We also thank David Rumelhart for useful discussions during the development of the basic architecture of TRACE and Eileen Conway, Mark Johnson, Dave Pare, and Paul Smith for their assistance in programing and graphics. Send requests for reprints to James L. McClelland, Department of Psychology, Carnegie-Mellon University, Schenley Park, Pittsburgh, PA 15213.

0010-0285/86 $7.50

the ability to trade cues off against each other in phoneme identification. At the word level, the model captures the major positive feature of Marslen-Wilson's COHORT model of speech perception, in that it shows immediate sensitivity to information favoring one word or set of words over others. At the same time, it overcomes a difficulty with the COHORT model: it can recover from underspecification or mispronunciation of a word's beginning. TRACE II also uses lexical information to segment a stream of speech into a sequence of words and to find word beginnings and endings, and it simulates a number of recent findings related to these points. The TRACE model has some limitations, but we believe it is a step toward a psychologically and computationally adequate model of the process of speech perception.

Consider the perception of the phoneme /g/ in the sentence "She received a valuable gift." There are a large number of cues in this sentence to the identity of this phoneme. 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 which 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 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 paper, we describe a model whose primary purpose is to account for the integration of multiple sources of information, or constraint, in speech perception. The model is constructed within a framework which appears to be ideal for the exploitation of simultaneous, and often mutual, constraints. This framework is the interactive activation framework (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1981, 1982). This approach grew out of a number of earlier ideas, some coming first from research on spoken language recognition (Marslen-Wilson & Welsh, 1978; Morton, 1969; Reddy, 1976) and others arising from more general considerations of interactive parallel processing (Anderson, 1977; Grossberg, 1978; McClelland, 1979).

According to the interactive-activation approach, information processing takes place through the excitatory and inhibitory interactions among a large number of processing elements called units. Each unit is a very simple processing device. It stands for a hypothesis about the input being processed. The activation of a unit is monotonically related

to the strength of the hypothesis for which the unit stands. Constraints among hypotheses are represented by connections. Units which are mutually consistent are mutually excitatory, and units that are mutually inconsistent are mutually inhibitory. Thus, the unit for /g/ has mutually excitatory connections with units for words containing /g/, and has mutually inhibitory connections with units for other phonemes. When the activation of a unit exceeds some threshold activation value, it begins to influence the activation of other units via its outgoing connections; the strength of these signals depends on the degree of the sender's activation. The state of the system at a given point in time represents the current status of the various possible hypotheses about the input; information processing amounts to the evolution of that state, over time. Throughout the course of processing, each unit is continually receiving input from other units, continually updating its activation on the basis of these inputs, and, if it is over threshold, it is continually sending excitatory and inhibitory signals to other units. This "interactive-activation" process allows each hypothesis both to constrain and be constrained by other mutually consistent or inconsistent hypotheses.

Criteria and Constraints on Model Development

There are generally two kinds of models of the speech perception process. One kind of model, which grows out of speech engineering and artifical intelligence, attempts to provide a machine solution to the problem of speech recognition. Examples of this kind of model are HEARSAY (Erman & Lesser, 1980; Reddy, Erman, Fennell, & Neely, 1973) HWIM (Wolf & Woods, 1978), HARPY (Lowerre, 1976), and LAFS/SCRIBER (Klatt, 1980). A second kind of model, growing out of experimental psychology, attempts to account for aspects of psychological data on the perception of speech. Examples of this class of models include Marslen-Wilson's COHORT Model (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978; Nusbaum & Slowiaczek, 1982); Massaro's feature integration model (Massaro, 1981; Massaro & Oden, 1980a, 1980b; Oden & Massaro, 1978); Cole and Jakimik's (1978, 1980) model of auditory word processing, and the model of auditory and phonetic memory espoused by Fujisaki and Kawashima (1968) and Pisoni (1973, 1975).

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

In extending the interactive activation approach to speech perception, we had essentially two questions: First, could the interactive-activation

approach contribute toward the development of a computationally sufficient framework for speech perception? Second, could it account for what is known about the psychology of speech perception? In short, we wanted to know, was the approach fruitful, both on computational and psychological grounds.

Two facts immediately became apparent. First, spoken language introduces many challenges that make it far from clear how well the interactive-activation approach will serve when extended from print to speech. Second, the approach itself is too broad to provide a concrete model, without further assumptions. Here we review several facts about speech that played a role in shaping the specific assumptions embodied in TRACE.

Some Important Facts about Speech

Our intention here is not to provide an extensive survey of the nature of speech and its perception, but rather to point to several fundamental aspects of speech that have played important roles in the development of the model we describe here. 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 difference between speech and print: speech is a signal which is extended in time, whereas print is a stimulus which is extended in space. The sequential nature of speech poses problems for a modeler, in that to account for context effects, one needs 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.

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 /Is/ (as in "kiss") or /Ift/ (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 not be a static representation, but should remain in a malleable form, subject to alteration as a result of influences arising from subsequent context.

Lack of boundaries and temporal overlap. A second 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 (Liberman, 1970; Fig. 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 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 stream 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 in running speech. 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 preceded 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

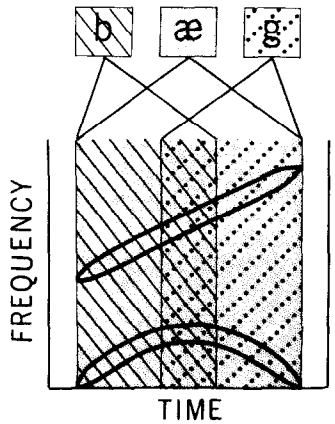

FIG. 1. A schematic spectrogram for the syllable "bag," indicating the overlap of the information specifying the different phonemes. Reprinted with permission from Liberman (1970).

word segmentation (Bond & Garnes, 1980), and certain segmentation decisions are easily influenced by contextual factors (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 third 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 Fig. 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). The extent of the restructuring depends on the unit selected and on the particular cue involved. But the problem is ubiquitous in speech.

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 (Miller, 1981), by morphological and prosodic factors such as position in 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, thereby effectively allowing the context to retune the perceptual mechanism on the fly.

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. Miller, Heise, & Lichten, 1951). This means that the speech mechanisms must be able to function, even with a highly degraded stimulus. In particular, as Thompson (1984), Norris (1982), and Grosjean and Gee (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 (Marslen-Wilson & 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 perception. 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, and Sejnowski, 1983; Barrow & Tenenbaum, 1978; Marr, 1982). Thus, though the model we present here is focussed on speech perception, we would hope that the ways in which it deals with the challenges posed by the speech signal are applicable in other domains.

The Importance of the Right Architecture

All four of the considerations listed above played an important role in the formulation of the TRACE model. The model is an instance of an interactive activation 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 four aspects of speech (see Elman & McClelland, 1984, for discussion). Thus, the TRACE model hinges as much on the particular processing architecture it proposes for speech perception as it does on the interactive activation processes that occur within this architecture.

Interactive-activation mechanisms are a class too broad to stand or fall on the merits of a single model. To the extent that computationally and psychologically adequate models can be built within the framework, the attractiveness of the framework as a whole is, of course, increased, but the adequacy of any particular model will generally depend on the particular assumptions that model embodies. It is no different with interactive-

activation models than with models in any other computational framework, such as expert systems or production systems.

THE TRACE MODEL

Overview

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 a hypothesis about a particular perceptual object occurring at a particular point in time defined relative to the beginning of the utterance.

A small subset of the units in TRACE II, the version of the model we focus on in this paper, is illustrated in Figs. 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 if it were a real speech stream, unfolding in time. Mock-speech inputs on the three illustrated dimensions for the phrase "tea cup" (/tikˆp/) are shown in Fig. 2. At any instant, input is arriving only at the units in one slice at the feature level. In terms of the display in Fig. 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 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 each of the three processing levels. This trace is unlike many traces,

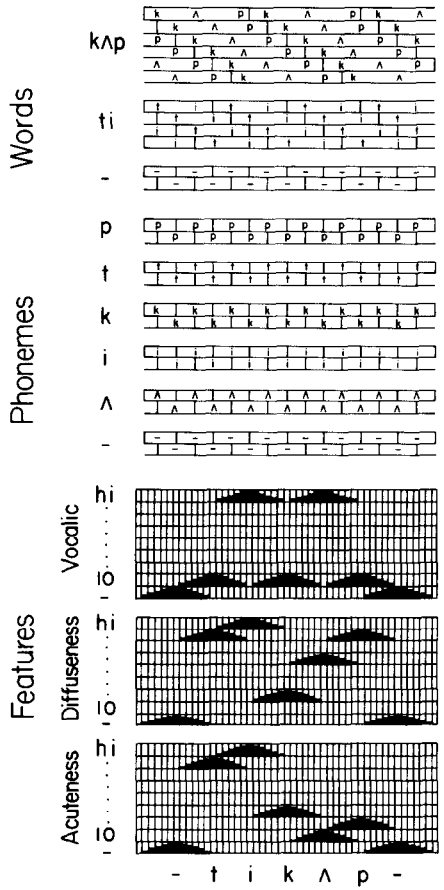

FIG. 2. A subset of the units in TRACE II. Each rectangle represents a different unit. The labels indicate the item for which the unit stands, and the horizontal edges of the rectangle indicate the portion of the Trace spanned by each unit. The input feature specifications for the phrase "tea cup," preceded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

though, in that it is dynamic, since it consists of activations of processing elements, and these processing elements continue to interact as time goes on. 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

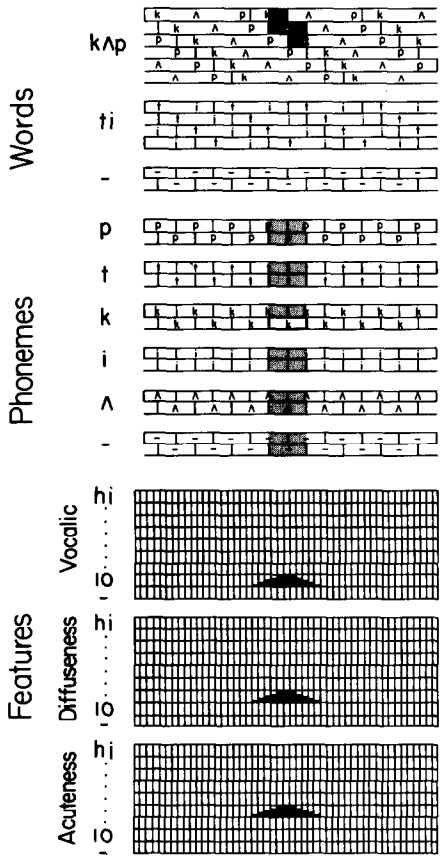

FIG. 3. The connections of the unit for the phoneme /k/, centered over Time Slice 24. The rectangle for this unit is highlighted with a bold outline. The /k/ unit has mutually excitatory connections to all the word- and feature-level units colored either partly or wholly in black. The more coloring on a units' 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.

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

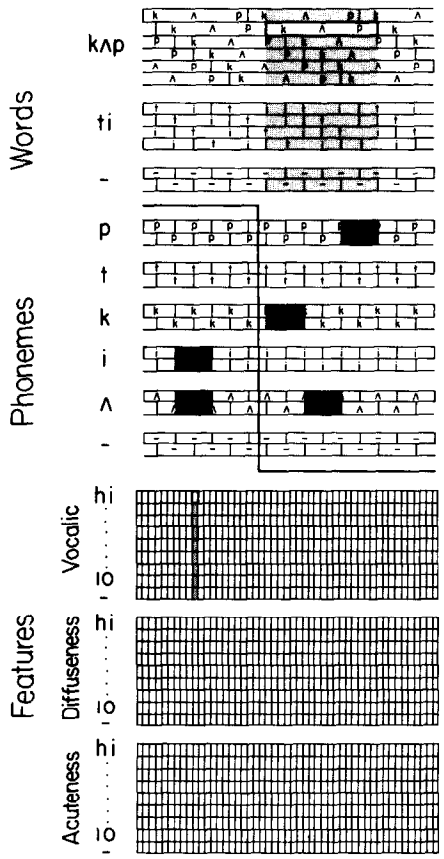

FIG. 4. The connections of the highlighted unit for the high value on the Vocalic feature dimension in Time Slice 9 and for the highlighted unit for the word /kˆp/ starting in Slice 24. Excitatory connections are represented in black, inhibitory connections in grey, as in Fig. 3.

level that are inconsistent have mutually inhibitory connections. All connections are bidirectional. Bidirectional excitatory and inhibitory connections of the unit for /k/ centered over Feature-slice 24 (counting from 0) are shown in Fig. 3; connections for the high value of the feature Vocalic in Slice 9 and for the word /kˆp/ with the /k/ centered over Slice 24 are shown in Fig. 4.

The interactive activation model of visual word recognition (McClelland & Rumelhart, 1981) included inhibitory connections between each unit on the feature level and letters that did not contain the feature, and between each letter unit and the words that did not contain the letter. Thus the units for *T* in the first letter position inhibited the units for all words that did not begin with *T*. However, more recent versions of the

visual model eliminate these between-level inhibitory connections, since these connections can interfere with successful use of partial information (McClelland, 1985; McClelland, 1986). Like these newer versions of the visual model, TRACE likewise contains no between-level inhibition. We will see that this feature of TRACE plays a very important role in its ability to simulate a number of empirical phenomena.

Sources of TRACE's architecture. The inspiration for the architecture of TRACE goes back to the HEARSAY Speech understanding system (Erman & Lesser, 1980; Reddy et al., 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 antonomous processes shared information.

The architecture of TRACE bears a strong resemblance to the "neural spectrogram" proposed by Crowder (1978, 1981) to account for interference effects between successive items in short-term memory. Like our Trace, Crowder's neural spectrogram provides a dynamic working memory representation of a spoken input. There are two important differences between the Trace and Crowder's neural spectrogram, however. First of all, the neural spectrogram was assumed only to represent the frequency spectrum of the speech wave over time; the Trace, on the other hand, represents the speech wave in terms of a large number of different feature dimensions, as well as in terms of the phonemes and words consistent with the pattern of activation at the feature level. In this regard TRACE might be seen as an extension of the neural spectrogram idea. The second difference is that Crowder postulates inhibitory interactions between detectors for spectral components spaced up to several hundred milliseconds apart. These inhibitory interactions extend considerably farther than those we have included in the feature level of the Trace. This difference does not reflect a disagreement with Crowder's assumptions. Though we have not found it necessary to adopt this assumption to account for the phenomena we focus on in this article, lateral extension of inhibition in the time domain might well allow the TRACE framework to incorporate many of the findings Crowder discusses in the two articles cited.

Context-Sensitive Tuning of Phoneme Units

The connections between the feature and phoneme level 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 adjusts 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 focus 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 were 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 were also able to demonstrate the efficacy of the idea of adjusting feature to phoneme connections on the basis of activations produced by surrounding context. With connection strength adjustment in place, the model was able to identify the stop consonant in 90% of a set of isolated monosyllables correctly, up from 79% with an invariant set of connections. This level of performance is comparable to what has been achieved by other machine-based phoneme identification schemes (e.g., Kopec, 1984) and illustrates the promise of the connection strength adjustment scheme for coping with variability due to local phonetic context. Ideas for extending the connection strength adjustment scheme to deal with the ways in which cues to phoneme identification vary with global variables (rate, speaker characteristics, etc.) are considered in the general discussion.

TRACE II, the version described in the present paper, was designed to account primarily for lexical influences on phoneme perception and

for what is known about on-line recognition of words, though we 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. Obviously, then, TRACE II sidesteps many fundamental issues about speech. 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.

The following sections describe TRACE II in more detail. First we consider the specifications of the mock-speech input to the model, and then we consider the units and connections that make up the Trace at each of the three levels.

Mock-Speech Inputs

The input to TRACE II was a series of specifications for inputs to units at the feature level, one for each 25-ms time slice of the mock utterance. These specifications were 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 consisted 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"). /^/ was also used to represent reduced vowels such as the second vowel in "target." There was also a "silence" segment represented by /–/. Special segments, such as a segment halfway between /b/ and /p/, were also used; their properties are described in descriptions of the relevant simulations.

A set of seven dimensions was used in TRACE II to represent the feature-level inputs. Five of the dimensions (Consonantal, Vocalic, Diffuseness, Acuteness, and Voicing) were taken from classical work in phonology (Jakobson, Fant, & Halle, 1952), though we treat each of these dimensions as continua, in the spirit of Oden and Massaro (1978), rather than as binary features. A sixth dimension, Power, was included because it has been found useful for phoneme identification in various machine systems (e.g., Reddy, 1976), and it was incorporated here to add an additional dimension to increase the differentiation of the vowels and consonants. The seventh dimension, the amplitude of the burst of noise that occurs at the beginning of word initial stops, was included to provide an additional basis for distinguishing the stop consonants, which otherwise differed from each other on only one or two dimensions. 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 was divided into eight value ranges. Each phoneme was assigned a value on each dimension; the values on the Vocalic, Diffuseness, and Acuteness dimensions for the phonemes in the utterance /tikˆp/ are shown in Fig. 2. The full set of values are shown in Table 1. Numbers in the cells of the table indicate which value on the indicated dimension was 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.

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

The correlations of the feature patterns for the 15 phonemes used are shown in Table 2. It is these correlations of the patterns assigned to the

TABLE 1
Phoneme Feature Values Used in TRACE II

Phoneme	Power	Vocalic	Diffuse	Acute	Cons.	Voiced	Burst
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	—

TABLE 2
Correlations of Feature Patterns of the Different Phonemes Used in Trace II

Phoneme	p	b	t	d	k	g	s	S	r	l	a	i	u	ˆ·
p	—	.76	.71	.56	.60	.46	.30							
b	.76	—	.56	.71	.46	.60								
t	.71	.56	—	.76	.56	.42	.35							
d	.56	.71	.76	—	.42	.56								
k	.60	.46	.56	.42	—	.77		.24						
g	.46	.60	.42	.56	.77	—								
s	.30		.35				—	.65						
S					.24		.65	—					.20	
r									—	.80	.29		.32	.37
l									.80	—	.32			.32
a									.29	.32	—	.65	.75	.67
i											.65	—	.65	.49
u								.20	.32		.75	.65	—	.59
ˆ·									.37	.32	.67	.49	.59	—

Note. Correlations of less than .20 have been replaced by blanks.

different phonemes, rather than the actual values assigned to particular phonemes or even the labels attached to the different mock-speech dimensions, that determine the behavior of the simulation model, since it is these correlations that determine how much an instance of one phoneme will tend to excite the detector for another.

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

The feature specification of each phoneme in the input stream extended over 11 time slices of the input. The strength of the pattern grew to a peak at the 6th slice and fell off again, as illustrated in Fig. 2. Peaks of successive phonemes were separated by 6 slices. Thus, specifications of successive phonemes overlapped, as they do in real speech (Fowler, 1984; Liberman, 1970).

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

Feature Level Units and Connections

The units at the feature level are detectors for features of the speech stream at particular moments in time. In TRACE II, there was a unit for each of the nine values on each of the seven dimensions in each time slice of the Trace. The figures show three sets of feature units in several time slices. Units for features on the same dimension within the same time slice are mutually inhibitory. Thus, the unit for the high value of the Vocalic dimension in Time Slice 9 inhibits the units for other values on the same dimension in the same time slice, as illustrated in Fig. 4. This figure also illustrates the mutually excitatory connections of this same feature unit with units at the phoneme level. In the next section we redescribe these connections from the point of view of the phoneme level.

The Phoneme Level and Feature–Phoneme Connections

At the phoneme level, there is a set of detectors for each of the 15 phonemes listed above. In addition, there is a set of detectors for the presence of silence. These 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 received excitatory input from feature units in a range of 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 Fig. 3, the strengths of the connections between the node for /k/ centered over Time Slice 24 and the nodes 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.

There are inhibitory connections between units at the phoneme level. Units inhibit each other to the extent that the speech objects they stand for represent alternative interpretations of the content of the speech stream at the same point in the utterance. Note that, although the feature specification of a phoneme is spread over a window of 11 slices, successive phonemes in the input have their centers 6 slices apart. Thus each phoneme-level unit is thought of as spanning 6 feature-level slices, as illustrated in Fig. 3. Each unit inhibits others in proportion to their overlap. Thus, a phoneme detector inhibits other phoneme detectors centered over the same slice twice as much as it inhibits detectors centered 3 slices away, and inhibits detectors centered 6 or more slices away not at all.

Word Units and Word–Phoneme Connections

There is a unit for every word in every time slice. Each of these units represents a different hypothesis about a word identity and starting location in the Trace. For example, the unit for the word /kˆp/ in Slice 24 (highlighted in Fig. 4) represents the hypothesis that the input contains the word "cup" starting in Slice 24. More exactly, it represents the hypothesis that the input contains the word "cup" with its first phoneme centered in Time Slice 24.

Word units receive excitation from the units for the phonemes they contain in a series of overlapping windows. Thus, the unit for "cup" in Time Slice 24 will receive excitation from /k/ in slices neighboring Slice

24, from /^/ in slices neighboring Slice 30, and from /p/ in slices neighboring Slice 36. As with the feature–phoneme connections, these connections are strongest at the center of the window and fall off linearly on either side.

The inhibitory connections at the word level are similar to those at the phoneme level. Again, the strength of the inhibition between two word units depends on the number of time slices in which they overlap. Thus, units representing alternative interpretations of the same stretch of phoneme units are strongly competitive, but units representing interpretations of nonoverlapping sequences of phonemes do not compete at all.

TRACE II has detectors for the 211 words found in a computerized phonetic word list that met all of the following constraints: (a) the word consisted only of the phonemes listed above; (b) it was not an inflection of some other word that could be made by adding "-ed," "-s," or "-ing"; (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 this criterion, 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 had detectors at the word level for silence (/–/), which was 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 bottom-up input is only coming into 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 cyclés has elapsed.

Details of Processing Dynamics

The interactive activation process in the Trace model follows the dynamic assumptions laid out in McClelland and Rumelhart (1981). Each unit has a resting activation value arbitrarily set at 0, a maximum activation value arbitrarily set at 1.0, and a minimum activation set at $-.3$. On every time cycle of processing, all the weighted excitatory and inhibitory signals impinging upon a unit are added together. The signal from one unit to another is just the extent to which its activation exceeds 0; if its activation is less than 0, the signal is 0.[1] Global level-specific excitatory, inhibitory, and decay parameters scale the relative magnitudes of different types of influences on the activation of each unit. Values for these parameters are given below.

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

Each processing time cycle corresponds to a single time slice at the feature level. This is actually a parameter of the model—there is no intrinsic reason why there should be a single cycle of the interactive-activation process synchronized with the arrival of each successive slice of the input. A higher rate of cycling would speed the percolation of effects of new input through the network relative to the rate of presentation.

Output Assumptions

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

[1] At the word level, the inhibitory signal from one word to another is just the square of the extent to which the sender's activation exceeds zero. This tends to smooth the effects of many units suddenly becoming slightly activated, and of course it also increases the dominance of one active word over many weakly activated ones.

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 criterial time during processing or (b) when a unit in the alternative set reaches a criterial strength relative to the activation of other alternative units. Once a decision has been made to make a response, one of the alternatives is chosen from the members of the set. The probability of choosing a particular alternative i is then given by the Luce (1959) choice rule:

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

when j indexes the members of the alternative set, and

$$S_i = e^{ka_i}.$$

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

Minimizing the Number of Parameters

At the expense of considerable realism, we have tried to keep TRACE II simple by using homogeneous parameters wherever possible. Thus, as already noted, the feature specifications of all phonemes were spread out over the same number of time slices, effectively giving all phonemes the same duration. The strength of the total excitation coming into a particular phoneme unit from the feature units was 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 were 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, the TRACE model still has a number of free parameters. These parameters are listed in Table 3. It should be noted that parameters are not in general directly comparable across levels. For example, phoneme-to-phoneme and word-to-word inhibition are not directly comparable to each other or to feature-to-phoneme inhibition, since feature-level units compete only within a single slice, while phoneme and word units compete in proportion to their overlap.

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 was remarkably robust under parameter variations, and no systematic search of the space of parameters was necessary. Generally, manipulations of parameters simply influence the magnitude or the timing of one effect or another without changing the basic nature of the effects observed. For example, stronger bottom-up excitation speeds things up and can indirectly influence the size of top-down effects, since, for example, stronger word level activations produce stronger feedback to the phoneme level. Stronger top-down excitation, of course, directly influences the magnitude of lexical effects. The one parameter that appeared to influence the qualitative behavior of the model was the strength of within-level inhibition. Stronger within-level inhibition make the model commit itself more strongly to slight early differences in activation among competing alternatives. There was, therefore, some tuning of this parameter to avoid early overcommitment that would prevent right context from exerting an influence under some circumstances. Finally, a low rate of feature-level decay was used to allow feature-level activations to persist after the input moved on to later slices.

The parameter values were held constant at the values shown in the

TABLE 3
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[a]	.04
Word-level inhibition[a]	.03
Feature-level decay	.01
Phoneme-level decay	.03
Word-level decay	.05

[a] Per three time-slices of overlap.

table throughout the simulations, except 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 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 .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 .03 and .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.

THE DYNAMICS OF PHONEME PERCEPTION

In the introduction, we motivated the approach taken in the TRACE model in general terms. In this section, we see that the simple concepts that lead to TRACE provide 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 can account for the time course of lexical influences on phoneme identification. What we hope to show here is that TRACE brings these phenomena, and several others not considered by either model, together into a coherent picture of the process of phoneme perception as it unfolds in time.

The present section consists of three main parts. The first focuses on lexical effects on phoneme identification and the conditions under which these effects are obtained. Here, we see how TRACE can account for the basic lexical effect, and we make it clear why lexical effects are only obtained under some conditions. 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 ef-

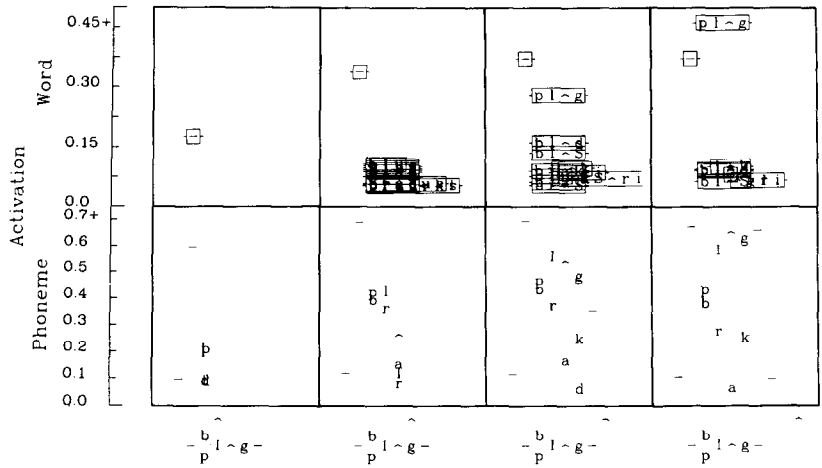

FIG. 5. Phoneme- and word-level activations at several points in the unfolding of a segment ambiguous between /b/ and /p/, followed by /l/, /ʌ/, and /g/. See text for a full explanation.

fects—namely, the contrasting phenomena of cue tradeoffs in phoneme perception and categorical perception. Here we see that TRACE provides an account of both effects as well as details of their time course. All three parts of this section illustrate how the simple mechanisms of mutual excitation and inhibition among the processing units of the Trace provide a natural way of accounting for the relevant phenomena. The section ends with a brief consideration of the ways in which TRACE might be extended to cope with several other aspects of phoneme identification and perception.

Lexical Effects

You can tell a phoneme by the company that it keeps.[2] In this section, we describe a simple simulation of the basic lexical effect on 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 approach 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

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

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 activations of these units generally get suppressed by the model, since the units on the peaks tend to dominate them). 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 /^/ 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 buildup of activation of the two interpretations of the first phoneme, followed by gradual buildups 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. In this case, it is only after the /g/ has come in that the model has information telling it whether the input is closer to "plug," "plus," "blush," or "blood" (TRACE's lexicon contains no other words beginning with /pl^/ or /bl^/). 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.

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

Figure 6 shows the temporal course of buildup of the strength of the

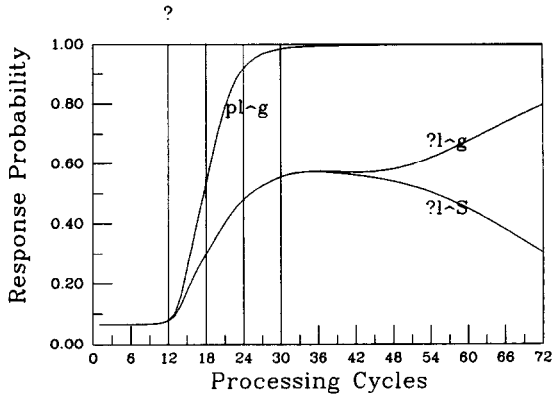

FIG. 6. The time course of the buildup in the strength of the /p/ response based on activations of phoneme units in Slice 12, in processing an ambiguous /b/–/p/ segment in /_l^g/, and the same segment in /–l^S/. The ambiguous segment is indicated by the "?". Also shown is the buildup of response strength for processing an unambiguous /p/ segment in /pl^g/. The vertical line topped with "?" indicates the point in time corresponding to the center of the initial segment in the input stream. Successive vertical lines indicate centers of successive phonemes.

/p/ response based on activations of the phoneme units in Slice 12 for two cases in which the initial segment is ambiguous between /p/ and /b/. In one case, the ambiguous segment is followed by /l^g/ (as in "plug"); in the other, it is followed by /l^S/ (as in "blush"). Given the model's restricted lexicon, which does not contain the word "plush," the lexical effect should lead to eventual dominance of the /p/ response in the first case, but a suppression of the /p/ response in the second case. The differences between the contexts do not begin to show up until after the center of the final phoneme, which occurs at Slice 30. The reason for this is simply that the information is not available until that point, because the phoneme that signals what the word will be comes at the very end of the word. The effect takes another few time slices to begin to influence the activation of the initial phoneme, because it percolates to the first phoneme by way of the feedback from the word or words that contain it.

Elimination of the lexical effect by time pressure. Fox (1982) has reported that the lexical effect on word initial segments is eliminated if subjects are given a deadline to respond within 500 ms of the ambiguous segment. Though they can correctly identify unambiguous segments in responses made before the deadline, these early responses show no sensitivity to the lexical status of the alternatives. Similar findings are also reported by Fox (1984).

Our model is completely consistent with Fox's results. Indeed, we have

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

But what about the fact that early responses to unambiguous segments can be accurate? TRACE accounts for this too. In Figure 7 we show the state of the Trace at various different points after the unambiguous /b/ in /bl^g/. Here, the /b/ dominates the /p/ from the earliest point. The analogous result is obtained, when the stimulus is /p/ in /pl^g/, and the activation for the initial phoneme is quite independent of whether or not the item is a word. The response strength for the case when /pl^g/ is presented in Fig. 6 shows that the probability of choosing /p/ is near unity within 12 processing cycles, or 300 ms of the initial segment, well before the deadline would be reached—and well before word identity specifying information is available.

Lexical effects late in a word. In the model, lexical effects on word-initial segments develop rather late, at least in the case where there is no context preceding the word. Of course, the exact timing of the development of any lexical effect would be dependent upon the set of words activated by the stimulus; if one word predominated early on, a lexical effect could develop rather earlier. In general, though, word-initial ambiguities will require time to resolve on the basis of lexical information.

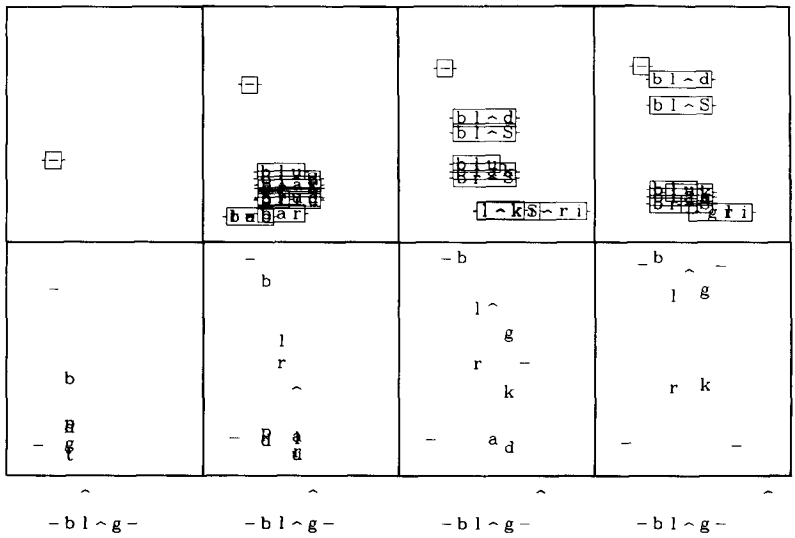

FIG. 7. The state of the Trace at various stages of processing the stream /bl^g/.

However, when the ambiguous segment comes late in the word, and the information that precedes the ambiguous segment has already established which of the two alternatives for the ambiguous segment is correct, TRACE shows a lexical effect that develops as the direct perceptual information relevant to the identity of the target segment is being processed. This phenomenon is illustrated in Fig. 8, which shows the state of the Trace at several points in time relative to an ambiguous final segment that could be a /t/ or a /d/, at the end of the context /targ^/. Within the duration of a single phoneme after the center of the ambiguous segment, /t/ already has an advantage over /d/. We therefore predict that Fox's results would come out differently, were he to use word-final, as opposed to word-initial, ambiguous segments. In such a case we would expect the lexical effect to show up well within the 500-ms deadline.

Dependence of the lexical effect on phonological ambiguity. One further aspect of the lexical effect that was noted by Ganong (1980) deserves comment. This is the fact that the lexical effect on the identity of a phoneme only occurs with segments which fall in the boundary region between two phonemes. For segments which are unambiguous examples of one category or the other, the effect is not obtained. TRACE is entirely consistent with this aspect of the data. The influence of the lexicon is simply another source of evidence, like that coming from the feature

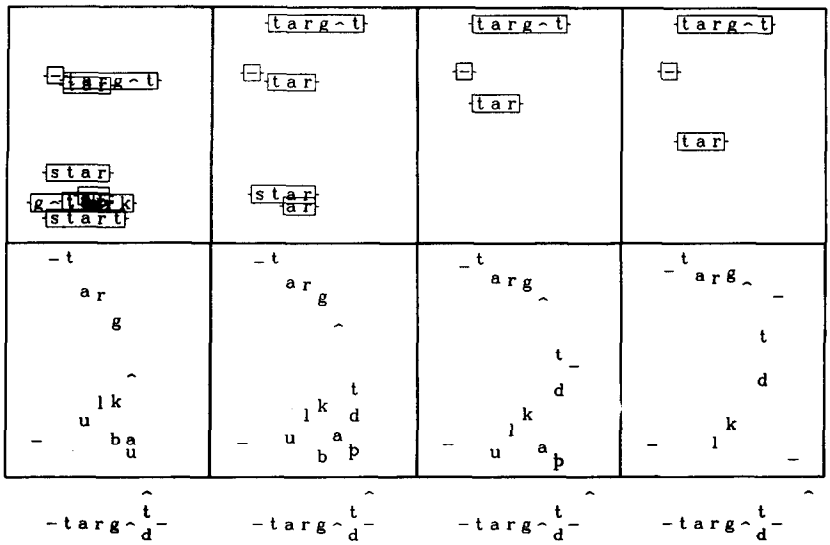

FIG. 8. The state of the Trace at several stages of processing the stream consisting of /targ^/ followed by a segment ambiguous between /t/ and /d/.

level, influencing the activation of one phoneme unit or another. When the bottom-up input is decisive, it can preempt any lexical bias effects. We have verified this in simulations presenting unambiguous tokens of /p/ or /b/, followed either by /lˆg/ or /lˆS/. In these simulations, the unit for the presented initial segment reaches a very high level of activation, independent of the following context. When the segment comes at the end of the word, the context exerts stronger effects, thus accounting for the fact that speech distortions are easier to detect when they come early in a word than when they come late (Marslen-Wilson & Welsh, 1978). However, even there, it is possible to override lexically based activations with clear bottom-up signals, although there may be some slowing of the activation process which would probably show up in reaction times.

It should be noted that TRACE's account of lexical effects is quite similar to the account offered by the feature integration theory of Massaro and Oden (1980a). Indeed, Massaro and Oden's model provides quantitative fits to Ganong's findings. We will make some mention of the slight differences in quantitative assumptions between the models below. For now, we note a more crucial difference: TRACE incorporates specific assumptions about the time course of processing which allows it to account for the conditions under which lexical effects will be obtained, as well as for the influence (or a lack thereof) of lexical effects on reaction times, to which we now turn.

Absence of lexical effect in some reaction-time studies. Foss and Blank (1980) presented some results which seemed to pose a challenge to interactive models of phoneme identification in speech perception. They gave subjects the task of listening to spoken sentences for occurrences of a particular phoneme in word-initial position. Reaction time to press a response key from the onset of the target phoneme was the dependent variable. In one example, the target was /g/ and the sentence was, *At the end of last year, the government. . . .* The subject's task was simply to press the response key upon hearing the /g/ at the beginning of the word *government.*

The principle finding of Foss and Blank's study was that it made no difference whether the target came at the beginning of a word or a nonword. Later studies by Foss and Gernsbacher (1983) indicate that other experiments which have found lexical or even semantic and syntactic context effects on monitoring latencies are flawed, and that monitoring times for word-initial phonemes are primarily influenced by acoustic factors affecting phoneme detectability, rather than lexical, semantic, or syntactic factors.

The conclusion that phoneme monitoring is unaffected by the lexical status of the target-bearing phoneme string seems at variance with the

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

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

The results of a simulation run illustrating these points are shown in Fig. 9. For this example, we imagine that the target is /t/. Note how during the initial syllable of both streams, little activation at the word level has been established. Even toward the end of the stream, where the information is just coming in which determines that "trugus" is not a word, there is little difference, because in both cases, there are several active word-level candidates, all supporting the word-initial /t/. It is only after the end of the stream that a real chance for a difference has occurred. Well before this time arrives, the subject will have made a response, since the strength of the /t/ response reaches a level sufficient to guarantee a high accuracy by about Cycle 30, well before the end of the word, as illustrated in Fig. 10.

Even though activations are quite rapid for unambiguous segments, these can still be influenced by lexical effects, provided that the lexical information is available in time. In Fig. 11, we illustrate this point for the phoneme /t/ in the streams /sikrˆt/ (the word "secret") and /gˆldˆt/ ("guldut," a nonword). The figure shows the strength of the /t/ response as a function of processing cycles, relative to all other responses based on activations of phoneme units at Cycle 42, the peak of the input specification for the /t/. Clearly, response strength grows faster for the /t/ in /sikrˆt/ than for the /t/ in /gˆldˆt/; picking an arbitrary threshold of .9 for response initiation, we find that the /t/ in /sikrˆt/ reaches criterion about 3 cycles or 75 ms sooner than the /t/ in /gˆldˆt/.

Studies showing lexical effects in reaction times. Marslen-Wilson (1980) has reported an experiment that demonstrates the existence of lexical effects in phoneme monitoring for phonemes coming at later points in words. For phonemes coming at the beginning of a word or at the end of the first syllable, he found no facilitation for phonemes in words rel-

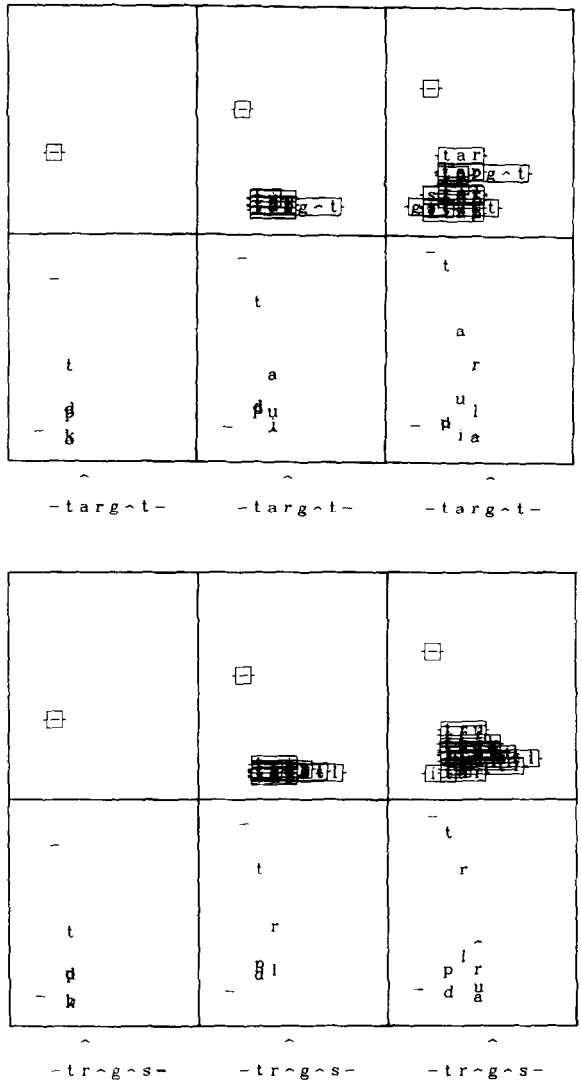

FIG. 9. State of the Trace at three different points during the processing of the word "target" (/targ^t/) and the nonword "trugus" (/tr^g^s/).

ative to phonemes in nonwords (in fact there was a nonword advantage for these early target conditions). For targets occurring at the end of the second syllable of a two-syllable word (like "secret"—though the stimuli in this particular experiment were Dutch) Marslen-Wilson found an 85-

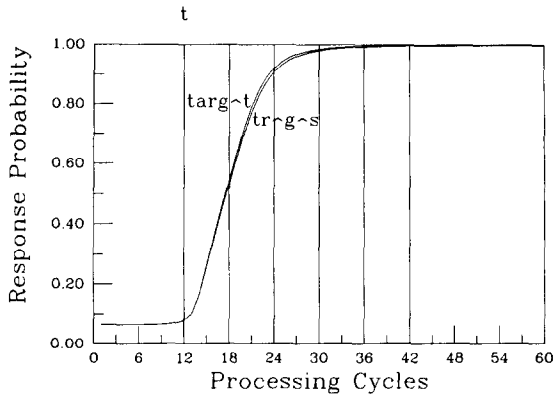

FIG. 10. Time course of growth in the probability of the /t/ response based on activations of phoneme units in Slice 12, during processing of /targ^t/ and /tr^g^s/. The vertical lines indicate the peaks on the feature patterns corresponding to the successive phonemes of the presented word.

ms advantage compared to corresponding positions in nonwords. This compares quite closely with the value of about 75 ms we obtained for the /sikr^t/–/g^ld^t/ example. At the ends of even longer words, the word advantage increased in size to 185 ms. Marslen-Wilson's result thus confirms that there are indeed lexical effects in phoneme monitoring—even for unambiguous inputs—but underscores the fact that there is no word advantage for phonemes whose processing can be completed long before lexical influences would have a chance to show up.

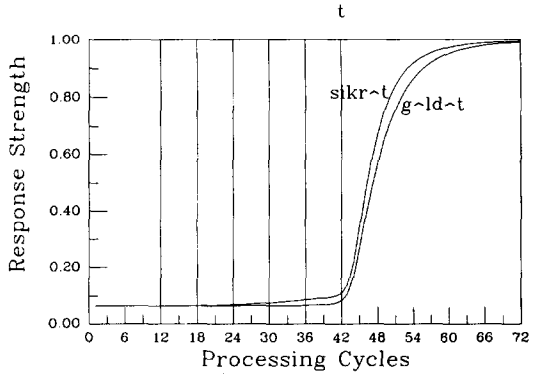

FIG. 11. Probability of the /t/ response as a function of processing cycles, based on activation of phoneme units at Cycle 42, for the stream /sikr^t/ ("secret") and /g^ld^t/ ("guldut"). Vertical lines indicate the peaks of the input patterns corresponding to the successive phonemes in either stream.

The TRACE model and Marslen-Wilson's COHORT model (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978) offer fairly similar interpretations of lexical effects in phoneme monitoring. Both models account for the growth in the effect as a function of position in the word. As in COHORT, lexical effects in TRACE depend on the point at which the pattern of activation at the word level begins to specify the identities of the phonemes. In COHORT, there is a discrete moment when this occurs—when the cohort of items consistent with the input is reduced to a single item. In TRACE, things are not quite so discrete. However, it will still generally be the case in TRACE that the size of the lexical effect will vary with the location of the "unique point," the point at which the bottom-up input remains consistent with only a single word. However, since Marslen-Wilson's experiments were performed with Dutch words, we have not been able to simulate his experimental demonstration of this effect in detail.

TRACE and COHORT make similar predictions in some situations, but not in all. In the next section, we consider a phenomenon which TRACE accounts for via the same mechanisms it uses to account for the lexical effects we have been considering. Here, the graded feedback from the word level to the phoneme level allows TRACE to account for an effect that would not be predicted by COHORT, unless additional assumptions were made.

Are Phonotactic Rule Effects the Result of a Conspiracy?

Recently, Massaro and 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 ony 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 Fig. 12.

The hypothesis that phonotactic rule effects are really based on word activations leads to a prediction: that we should be able to reverse these effects if we present items that are supported strongly by one or more lexical items even if they violate phonotactic rules. A recent experiment by Elman (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

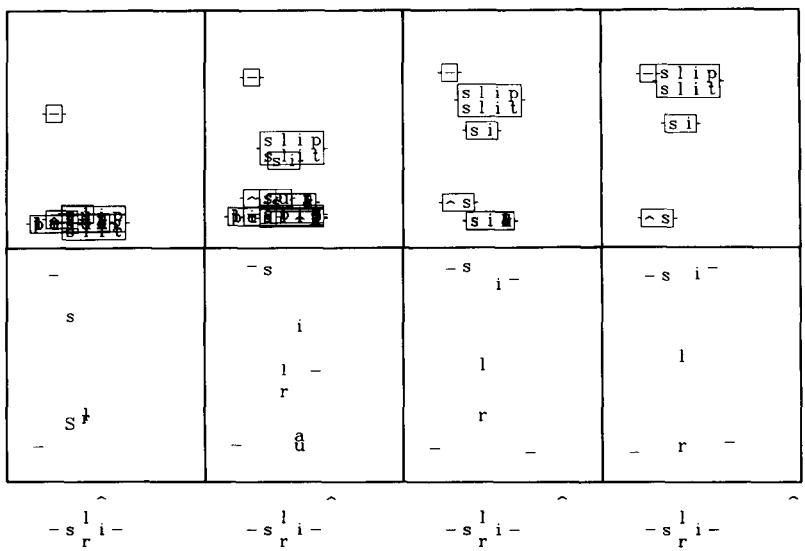

FIG. 12. State of the Trace at several points in processing a segment ambiguous between /l/ and /r/, in the context /s_i/. The units for "sleep" (/slip/) and "sleet" (/slit/) are boxed together since they take on identical activation values.

was particularly close to a word ("bwiffle"–"dwiffle"). Results of the experiment are shown in Table 4. 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 subject's 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.

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 Fig. 13. 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.

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

TABLE 4
Percentage Choice of Phonotactically Irregular Consonant

Stimulus type	Example	Percentage of identifications as "illegal" phoneme[a]
Legal word/illegal nonword	dwindle/bwindle	37
Legal nonword/illegal nonword	dwiffle/bwiffle	46
Legal nonword/illegal nearword	dwacelet/bwacelet	55

[a] $F(2,34) = 26.414$, $p < .001$.

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 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—specifically, the ability to trade one feature of a phoneme off against another—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-

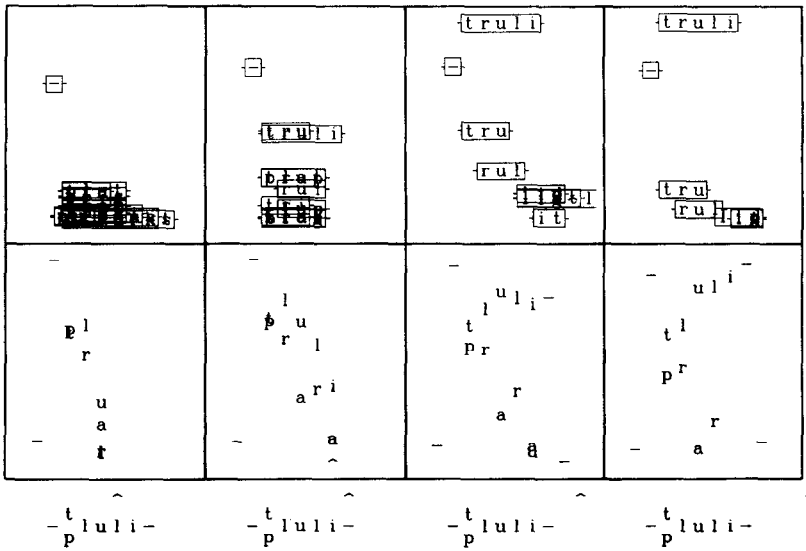

FIG. 13. State of the Trace at several points in processing an ambiguous /p/–/t/ segment followed by /luli/

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: (1) the /a/ and /–/ had no 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; (2) /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 14 illustrates how F1OF tends to increase as voice onset time is delayed.

Summerfield and Haggard (1977) have shown that subjects are sensitive both to VOT and to F1OF and that it is possible to trade one of these cues off against the other. Thus, the boundary between /ga/ and /ka/ shifts to longer VOTs when F1 starts off lower rather than higher.

Categorical perception and trading relations among cues have been studied on a variety of different continua by a variety of different investigators. We have chosen to focus on the VOT and F1OF features, as exemplified by the /ga/–/ka/ continuum, because there is data on tradeoffs between these cues (Summerfield & Haggard, 1977), and because

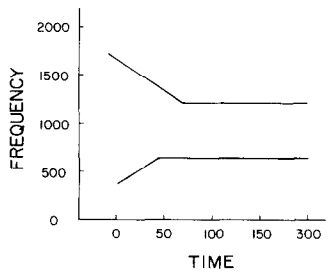

FIG. 14. Schematic diagram of a syllable that will be heard as /ga/ or /ka/, depending on the point in the syllable at which voicing begins. Prior to the onset of voicing, F2 (top curve) is energized by aperiodic noise sources, and F1 is "cut back" (the noise source has little or no energy in this range). 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's (1982) analysis of the parameters of formants in natural speech, and are similar to those used in many perceptual experiments.

several categorical perception studies of VOT continua (using /g/–/k/, /d/–/t/, or /b/–/p/ stimuli) have covaried both VOT and F1OF, if only because F1OF tends to covary with VOT when realistic stimuli are used (e.g., Pisoni & Lazarus, 1974; Samuel, 1977). Though the simulations use a /g/–/k/ continuum, we consider several categorical perception experiments using /d/–/t/ and /b/–/p/ continua, since the same dimensions can differentiate the two members of both of these other pairs. We also consider data obtained in experiments on other continua, using other cues. We could easily have repeated the simulations with other sets of continua; however, the general qualitative form of the results would be the same. What would vary from case to case would be the magnitude of the effect of a step along a given dimension.

The pattern of excitatory input to the VOT and F1OF detectors produced by the canonical mock speech /g/ and /k/ used in the simulations are illustrated in Fig. 15.

Trading relations. TRACE quite naturally tends to produce trading relations between features, since it relies on the weighted sum of the excitatory inputs to determine how strongly the input will activate a particular phoneme 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 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 Fig. 16 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, Figs. 4 and 5), though the transitions are often somewhat steeper (see below for a discussion of the issue of steepness). We have chosen to present the shallower curves reported by Denes because in them we see clearly that there are cases in which a cue that favors one of the two phonemes to a moderate degree will give rise to the perception of the other phoneme when paired up with a strong cue that favors the other

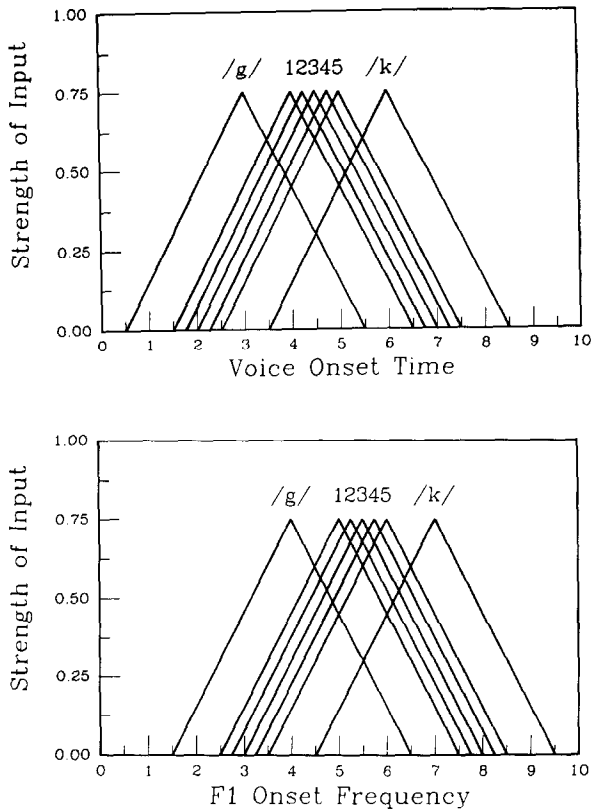

FIG. 15. Canonical feature-level input for /g/ and /k/, on the two dimensions that distinguish them, and the patterns used for the five intermediate values used in the trading relations simulation. Along the abscissa of each dimension the nine units for the nine different value ranges of the dimension 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. That is, the canonical input pattern for each phoneme exactly matches the strengths of the corresponding feature–phoneme connections. Numbered curves on each dimension show the feature patterns used in the trading relations simulation.

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

To see if TRACE would simulate the basic trade-off effect obtained by Summerfield and Haggard, and to see if it would produce the same shape

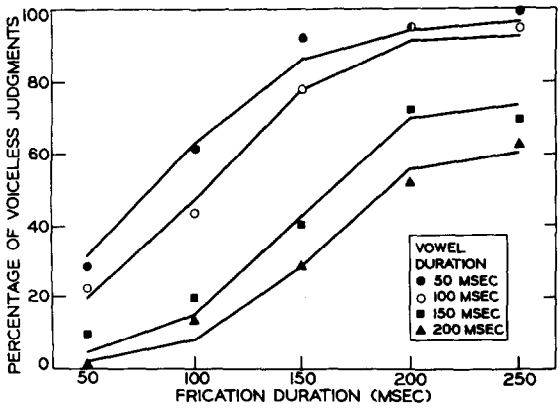

FIG. 16. Results of an experiment demonstrating the trade-off between two cues to the identity of /s/ and /z/. Data from Denes, 1955, fitted by the model of Massaro and Cohen, 1977. ●, 50 ms; ○, 100 ms; ■, 150 ms; ▲, 200 ms. Reprinted with permission from Massaro and Cohen (1977).

trade-off curves as have been generally reported, 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 Fig. 15, 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–/. In this and all subsequent simulations we report in this paper, the peak of the initial silence phoneme occurred at Time Slice 6 in the input, and the peaks of successive phoneme segments occurred at six slice intervals. Thus, for these stimuli, the peak on the intermediate phonetic segment occurred at Slice 12, the peak of the following vowel occurred at Slice 18, and the peak of the final silence occurred at Slice 24. For each input presented, the interactive activation process was allowed to continue through a total of 60 time slices, well past the end of the input. The state of the Trace at various points in processing, for the most /g/-like of the 25 stimuli, is shown in Fig. 17. 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 qualitative appearance of the results if a different decision time is used; earlier decision times are associated with smaller differences in relative activation between the /g/ and /k/ phoneme units, and later ones with larger differences, but the general pattern is the same.)

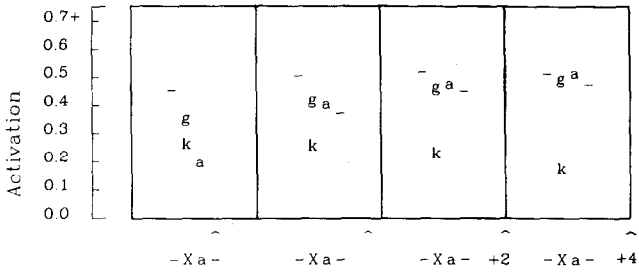

FIG. 17. The state of the Trace at various points during and after the presentation of a syllable consisting of the most /g/-like of the 25 intermediate segments used in the trading relations experiment, represented by /X/, preceded by silence and followed by /a/, then another silence.

Response probabilities were computed using the formulas given earlier for converting activations to response strengths and strengths into probabilities. The resulting response probabilities, for each of the 25 conditions of the experiment, are shown in Fig. 18. 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, but 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. For example, there are three different combinations of feature values which lead to a probability between .82 and .85 of choosing /k/: (1) the neutral value of the VOT dimension coupled with the most /k/-like value on the F1OF dimension; (2) the neutral value on the F1OF dimension coupled with the most /k/-like value of the VOT dimension; and (3) the somewhat

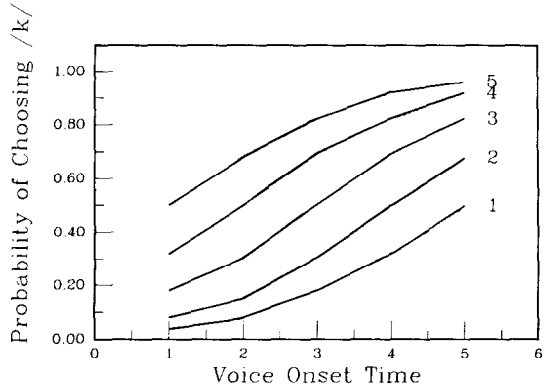

FIG. 18. Simulated probability of choosing /k/ at Time Slice 60. for each of the 25 stimuli used in the trading relations simulation experiment. Numbers next to each curve refer to the intermediate pattern on the F1OF continuum used in the five stimuli contributing to each curve. Higher numbers correspond to higher values of F1OF.

/k/-like values on both dimensions. In terms of Summerfield and Haggard's measure, the value of VOT needed to achieve 50% probability of reporting /k/, we can see that the VOT needed increases as the F1OF decreases, just as these investigators found.

Cue trade-offs in phoneme identification are accounted for in detail by the feature integration model of Oden and Massaro (1978; Massaro, 1981; Massaro and Oden, 1980a, 1980b). While we have shown how TRACE can account for the basic trade-off effect and the general form of the trade-off curves, we have not yet attempted the kinds of detailed fits that Massaro, Oden, and collaborators have reported in a number of studies. However, the models are quite similar, so it seems rather unlikely that cue trade-off data would be able to discriminate between them. And both make special assumptions about lack of invariance of cues to phoneme identity across contexts.

One apparent dissimilarity between the models deserves comment. Whereas cue strengths are combined multiplicatively in the determination of response strengths in the feature integration model, they are combined additively in the bottom-up inputs to the units in TRACE. However, in TRACE, two further computational steps take place before these inputs result in response strengths. First, the interactive-activation process enhances differences between competing units. Second, the resulting unit activations are subjected to an exponential transformation. Just this second step by itself would transform influences that have additive effects on unit activations into influences that have multiplicative effects on response strength. Thus, the models would be mathematically equivalent if the interactive activation process were simply replaced by a linear, additive combination of inputs to the units. In quantitative formulations of the interactive activation process closely related to the ones we use (Grossberg, 1978), what the interactive activation process does is simply rescale the unit activations, preserving the ratios of their bottom-up activation but keeping them bounded. Though our version of these equations does not do this exactly, the ways in which it deviates from this would be difficult to use as the basis for an empirical distinction between the TRACE approach and the feature integration model. Thus, up to a point, we can see TRACE as (approximately) implementing the computations specified in Oden and Massaro's model. The models differ, though, in that TRACE is dynamic and in that it incorporates feedback to the phoneme level. This allows TRACE to account for categorical perception in a different way.

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 F1OF 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 F1OF, 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 F1OF dimensions. The locations of the peak activation values on each of these two continua are shown in Fig. 19.

Figure 20 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 excitation 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 5), sends 88% as much activation to /g/ as to /k/.

The figure also indicates, in the second panel, the resulting activations

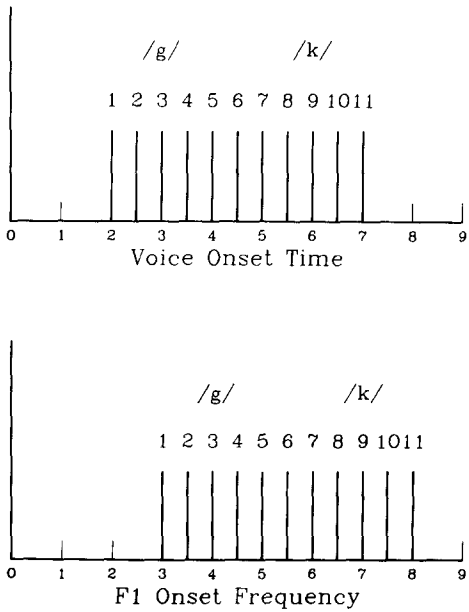

FIG. 19. Locations of peak activations along the VOT and F1OF dimensions, for each of the 11 stimuli used in the categorical perception simulation.

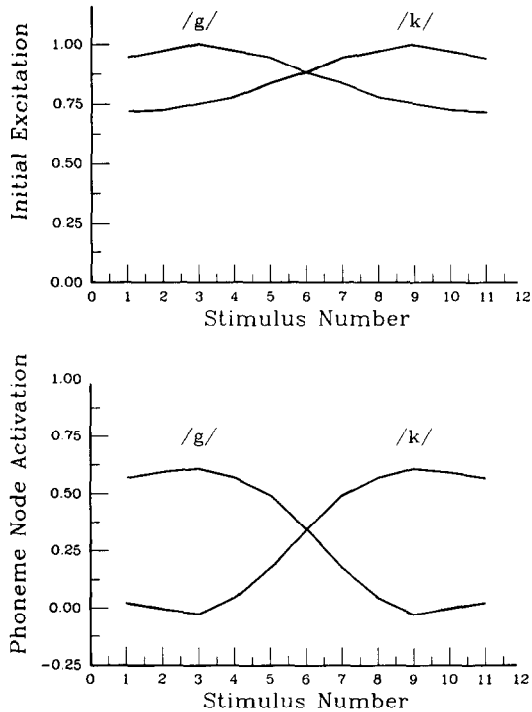

FIG. 20. Effects of competition on phoneme activations. The first panel shows relative amounts of bottom-up excitatory input to /g/ and /k/ produced by each of the 11 stimuli used in the categorical perception simulation. The second panel shows the activations of units for /g/ and /k/ at Time Cycle 60. Stimuli 3 and 9 correspond to the canonical /g/ and /k/, respectively.

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. Once one phoneme is slightly more strongly activated than the other, it exerts a stronger inhibitory influence on the other than the other can exert on it. The net result is that "the rich get richer." This general property of competitive inhibition mechanisms was discussed by McClelland and Rumelhart (1981), following earlier observations by Grossberg (see Grossberg, 1978, for a discussion) and Levin (1976); it is also well known as one possible basis of edge enhancement effects in low levels of visual

information processing. A second cause of the sharpening of the activation curves is the phoneme-to-feature feedback, which we 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 Fig. 20 are shown in Fig. 21 along with the *1BX* discrimination function, which is discussed below. 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 0 activation. Of course, the amount of separation in the activations that is necessary for any given level of performance is a matter of parameters; the relevant parameter here is the scale factor used in the exponential transformation of activations. The value used for this parameter in the present simulations (10) was the same as that used in all other cases where we translate activation into response probability, including the trading relations simulation.

Some readers may be puzzled as to why TRACE II exhibits a sharp identification function in the categorical perception experiment, but shows a much more gradual transition between /g/ and /k/ in the trading relations simulation. The reason is simply that finer steps along the VOT and F1OF continua were used in the trading relations simulation. All of the stimuli for the trading relations simulation lie between Stimuli 6 and 4 in the categorical perception simulation.

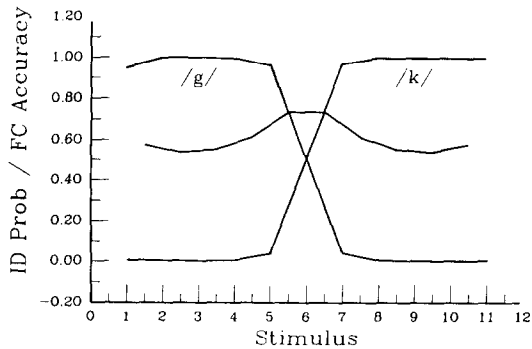

FIG. 21. Simulated identification functions and forced-choice accuracy in the *ABX* task.

This obviously brings out the fact that the apparent steepness of the identification function depends on the grain of the sampling of different points along the continuum between two stimuli, as well as a host of other factors (Lane, 1965). Whether an empirical or simulated identification function looks steep or not depends on the selection of stimuli by the experimenter or modeler. However, it is worth noting that the steepness of the identification function is independent of the presence of trading relations, at least in the simulation model. That is, if we had used more widely separated steps along the VOT and F1OF dimension, we would have obtained much steeper identification functions. The additivity of excitatory inputs would still apply, and thus it would still be possible to trade cues off against each other.

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 criterial level, we would find that subjects 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).

The sharpening the model imposes on the identification function, in conjunction with the fact that it can trade one feature off against another, shows how the model, like human perceivers of speech, can be both flexible and decisive at the same time. These aspects of TRACE are shared with the feature integration model (Massaro, 1981). However, the TRACE model's decisiveness extends even further than we have observed thus far; feedback from the phoneme to the feature level tends to cause the model to obliterate the differences between input feature patterns that result in the identification of the same phoneme, thus allowing the model to provide an account not only for sharp identification functions, but also for the fact that discriminability of speech sounds is far poorer within categories than it is between categories.

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

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

The way it works is this. When a feature pattern comes in, it sends more excitation to some phoneme 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.

An impression of the magnitude of this effect is illustrated in Fig. 22, which shows how different the feature patterns of adjacent stimuli 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. Furthermore, the correlation considers only the feature pattern on the feature

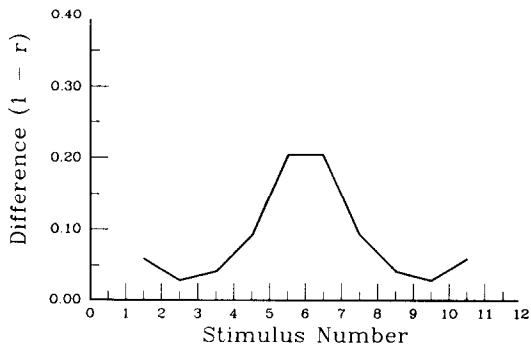

FIG. 22. Differences between patterns of activation at the feature level at Cycle 60, for pairs of stimuli one step apart along the /g/–/k/ continuum used for producing the identification functions shown previously in Fig. 21. The difference measure is the correlation of the two patterns, subtracted from 1.0; thus, if the two patterns correlated perfectly, their difference would be 0.

units in Time Slice 12, right at the center of the input specification. If all dimensions are considered, the values of the difference measure are reduced overall, but the pattern is the same. Inclusion of feature patterns from surrounding slices likewise makes little difference.

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

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

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

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

and similarly for S_{bx}. (The exponential transformation is required to translate correlations, ranging from $+1$ to -1, into positive values, so that Luce's ratio rule can be used. The same transformation is used for translating activations into response strengths in identification tasks.) Here k_r is the parameter that scales the relation between correlations and strengths. These assumptions are consistent with the choice assumptions made for identification responses. The resulting response probabilities, for one choice of the parameter k_r (5) are shown in Fig. 21 (the exponentiation parameter k_r is different than the parameter k used in generating identification probabilities from activations because correlations and activations are not on equivalent scales).

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

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 which 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 et al., 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, which has had considerable success accounting for categorical perception data, assumes that phoneme identification can be based either on precategorical information or on the results of the phoneme identification process. Since it is assumed that feature information decays rapidly (especially for consonant features—see below), responses must often be based solely on the output of the phoneme identification process, which is assumed to provide a discrete code of the sequence of phonemes. This interpretation accounts for much of the data on categorical perception quite well. Indeed, it is fairly difficult to find ways of distinguishing between a feedback model and one that attributes categorical perception to a loss of information from the feature level coupled with a reliance on a more abstract code. Both feedback models and dual code models can 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 one way. The feedback account seems to differ most clearly from a limited feature access account in its 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), simply in terms of the fact 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 dif-

ferent 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. For example, Stimulus 1 in our simulations falls below the canonical /g/ stimulus, and therefore activates the /g/ phoneme detector less strongly than stimuli closer to the canonical /g/. A similar thing happens with the /k/. This results in less "canonicalization" of the extreme stimuli, and produces a "W"-shaped discrimination function, as shown in Fig. 22.

There is some evidence bearing on this aspect of TRACE's account of categorical perception. Samuel (1977) has reported *ABX* discrimination data that show noticeable minima in the discrimination function near the canonical stimuli within each category on a /d/–/t/ continuum. Indeed, Samuel's account of this effect, though not couched in terms of interactive activation processes, has a great deal of similarity to what we see in TRACE; he suggests that near-canonical items are more strongly assimilated to the canonical pattern. Unfortunately the effect we seek is fairly subtle, and so it will be difficult to separate from noise. In Samuel's experiment, the effect is fairly clear-cut at both extremes of the VOT continuum in three observers at the end of extensive training, as shown in Fig. 23, and even unpracticed subjects tend to show the effect toward the high end of the VOT continuum, well past the prototype for /t/.

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 readout 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.

Other Phenomena at the Phoneme Level

The literature on phoneme perception includes several further findings we have not yet been able to consider in detail. The next few paragraphs consider one of these findings and how it might be accommodated in the TRACE model.

Effects of global and local context on phoneme identification. In our simulations of trading relations, we have shown that the criterial value needed on one dimension of stimulus variation can be affected by other dimensions. Thus, when the onset of F1 is relatively high, shorter voicing latencies are needed to perceive a sound as unvoiced. Other factors also influence the phoneme perceived as a result of a particular featural input.

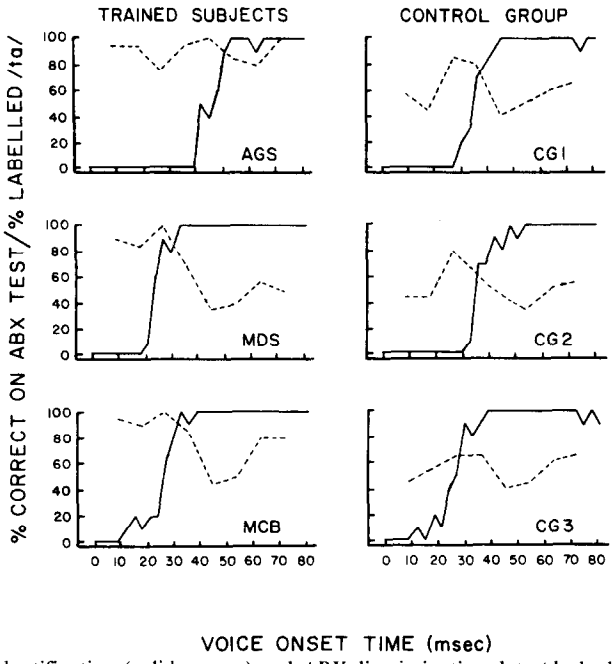

FIG. 23. Identification (solid curves) and *ABX* discrimination data (dashed curves) from three practiced and three naive subjects. Simplified and reprinted, with permission, from Samuel (1977).

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 discussion of all of these sorts of influences on the boundaries between phonemes.

It has been suggested by Miller, Green, and Schermer (1984) and by Repp and Liberman (1984) that these different effects may have different sources. In particular, Miller et al. (1984) suggest that lexical effects and semantic and syntactic influences on the one hand may be due to a different mechanism than influences such as speech rate and coarticulatory influences due to local phonetic context.

The assumptions we have incorporated into TRACE make a similar distinction. In TRACE I, we have accounted for effects of phonetic context by allowing activations of units to influence the feature-to-phoneme connections in adjacent time slices (see Elman & McClelland, in press, for details). In the discussion, we consider ways of extending the connection modulation idea to accommodate effects of variations in rate and

speaker parameters. Our main point here is that connection modulation is quite a different mechanism than the simple additive combination of excitatory influences that underlies the way TRACE accounts for trade-offs among the cues to a single phoneme or for the effects of top-down influences on the phoneme boundary.

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, both in reaction time and in response choice measures; "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. 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 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 mid 1970s, 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 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; Cole and Jakimik, 1978, 1980; Marslen-Wilson, 1980; Marslen-Wilson and 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 ms (N = 30 to 50 in different studies). Later presentations are successively lengthened in N-ms 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 exepected 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 (Salasso & 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 what 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 (e.g., one or two features) deviations 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 which 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 which 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 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 deliniated 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 dominate all others, but if no word matches perfectly, a word that provides a close fit to the phoneme sequence can eventaully win out over words that provide less adequate matches. The exact metric of "closeness of fit" depends, of course, on a large number of details. In the absence of such a metric, a simple count of the number of acoustic features differing between a lexical item and a presented stimulus can provide a useful first approximation, but other factors such as stress, location of differences within the word, and discriminability of the differing features will of course come into play.

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 Fig. 24. 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 and losing, but rather the fact that both tend to stay in the game. Such neologisms can suggest a poetic

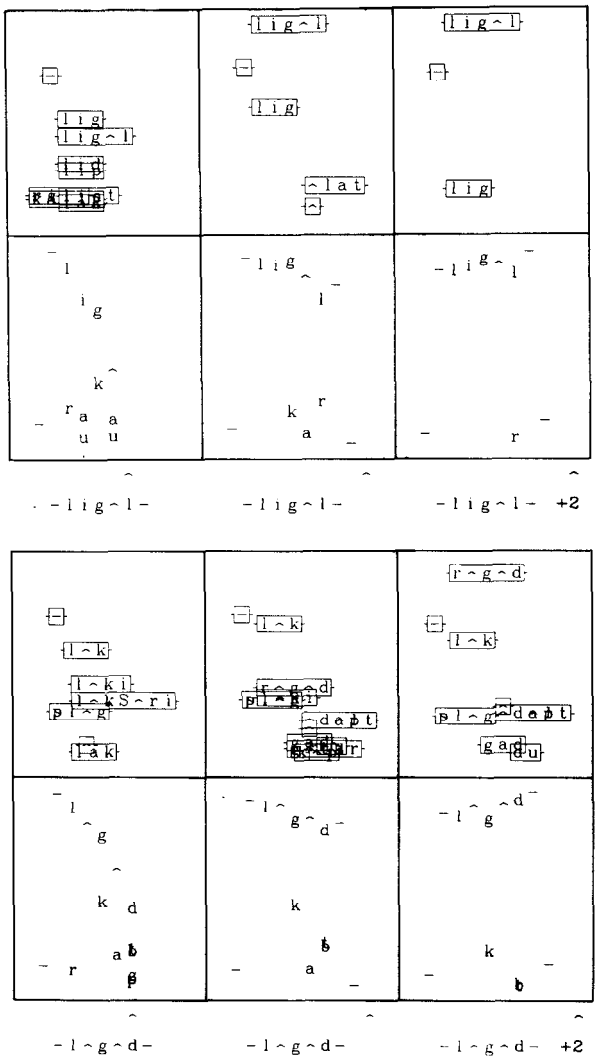

FIG. 24. State of the Trace at two points during processing of "legal" and "lugged."

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 dificulty 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 25 shows the state of the Trace at various points in processing this word, and Fig. 26 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 this figure, the response strength of "product" is simply set to 1.0 at each time slice and the response strengths of units for other words are plotted in terms of the ratio of their strength, divided by 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 to the simulation program's stressless encoding of them!) in the 1st, 2nd, 3d, 4th, and 5th phonemes, respectively. Figure 26 shows that these items begin to drop out of "contention" 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

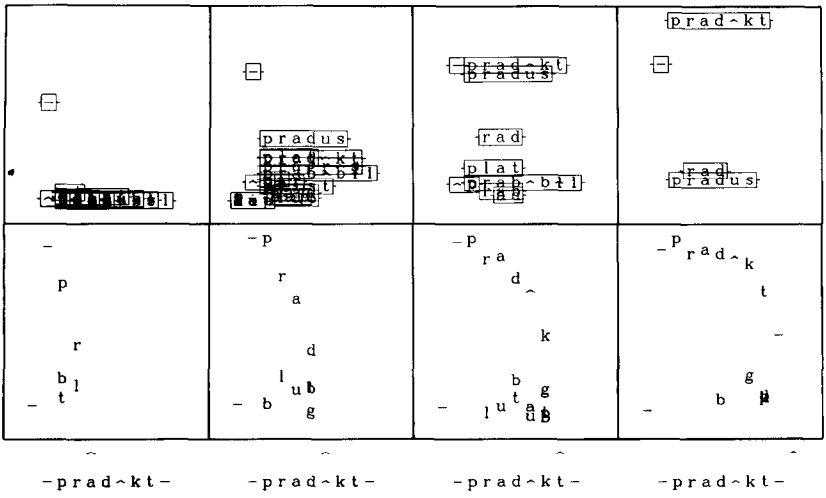

FIG. 25. State of the Trace at various points in processing the word "product" (/prad^kt/).

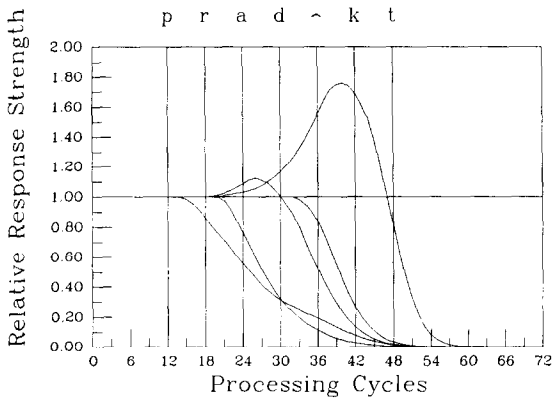

FIG. 26. Response strengths of the units for several words relative to the response strength of the unit for "product" (/prad^kt/), as a function of time relative to the peak of the first phoneme that fails to match the word. The successive curves coming off of the horizontal line representing the normalized 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.

produce in this case, though of course the drop-off would be assumed to be an abrupt, discrete event.[3]

There is one aspect of TRACE's behavior which differs from that of 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").

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 when many candidates are active, and so they tend to suffer in comparison to short words as a result.

[3] 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 Fig. 26 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.

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. It 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 well 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 which 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 (Cole & Jakimik, 1980; Grosjean & Gee, 1984; Thompson, 1984), they are not always sufficient, particularly in fluent speech. It would thus 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 be joint results of the mechanisms of word recognition.

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 inhibits 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 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 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 would 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 Fig. 27. 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

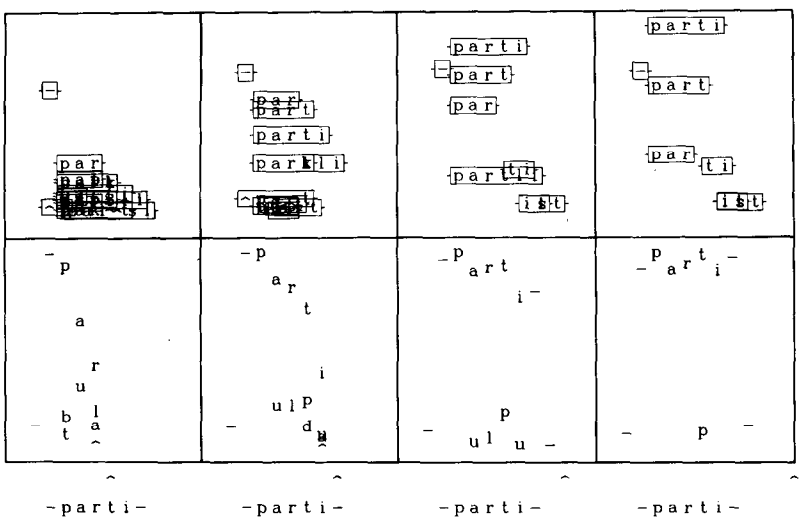

FIG. 27. The state of the Trace at various points during processing of /parti/.

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 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, as illustrated in Fig. 27. This behavior of the model leads to the prediction that short words embedded 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 Fig. 28, 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 Fig. 28. 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"

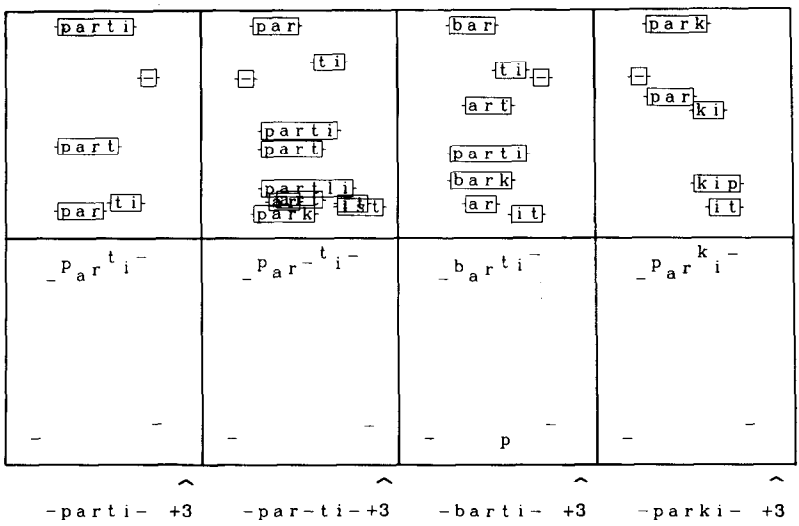

FIG. 28. State of the Trace after processing the streams /parti/, /par-ti/, /barti/, and /parki/.

and "tea" do not overlap. Thus, "art" receives inhibition from both "bar" 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 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 that 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 Fig. 28.

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 Fig. 29. In this case, higher levels could provide an additional source of information that would help the model choose between overlapping and nonoverlapping interpretations.

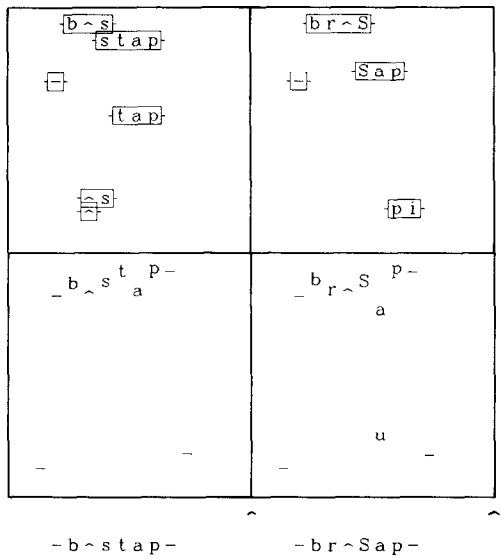

FIG. 29. State of the Trace at the end of the streams /bustap/ ("bus stop" or "bus top") and /bruSap/ ("brush shop").

The simulations we have reported show that the word activation/competition mechanism can go a long way toward providing a complete interpretation of the input stream as a sequence of words. As a word is beginning to come in, the model tends to prefer shorter words consistent with the input stream over longer ones. As the input unfolds through time, however, the model tends to prefer to interpret streams of phonemes as single longer words rather than as a sequence of short words; and it tends to find parses that account for each phoneme once. But it does not insist upon this, and will occasionally produce an interpretation that leaves part of the stream of phonemes unaccounted for or which accounts for part of the stream of phonemes twice. Often enough, it will also leave an alternative to its "preferred parse" in a strong position, so that both the preferred parse and the alternative would be available to higher levels and subject to possible reinforcement by them.

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., 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 like the following:

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

than they were when the word preceding the target (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 Fig. 30. 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, the subject would have to wait until he is well into the word "target" before it becomes clear that the first /t/ in target is in fact a word-initial /t/.

In reality, the situation is probably not as bleak for the perceiver as it appears in this example, because in 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

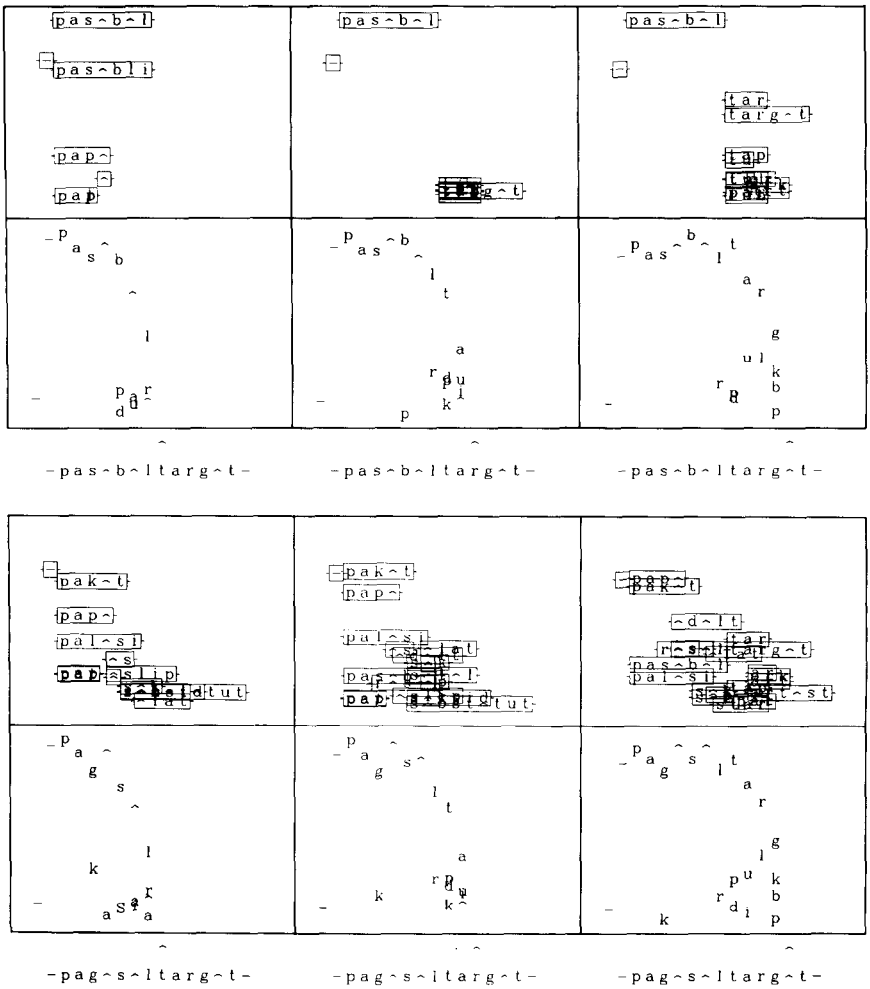

FIG. 30. State of the Trace at several points during the processing of "possible target" and "pagusle target."

lexical status of one part of a speech stream plays an important role in determining where the beginning of the next word must be.

The long and short of word identification. One 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 the preferred hypothesis at this point, we do not have enough information to say unequivocally that the word "tar" will not turn out to be "target"

or "tarnished" or one of several other possibilities. It is only after 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 Fig. 31. 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.

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

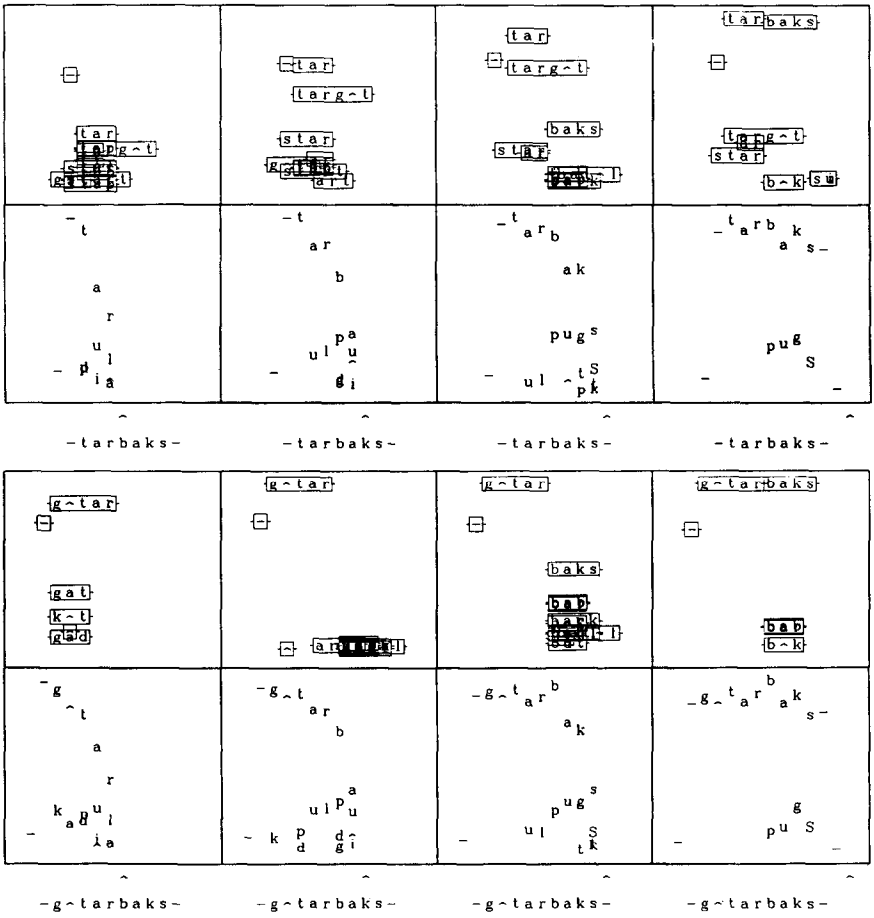

FIG. 31. State of the Trace at several points in processing "tar box" and "guitar box."

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" /gˆtarbaks/. 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 the 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 Fig. 32. The figure shows the state of the Trace at several points during the processing of the stream /SiSˆtˆbaks/. By the end, the words of the phrase "She shut a box," which fits the input perfectly with no overlap, dominate all others.

This 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.

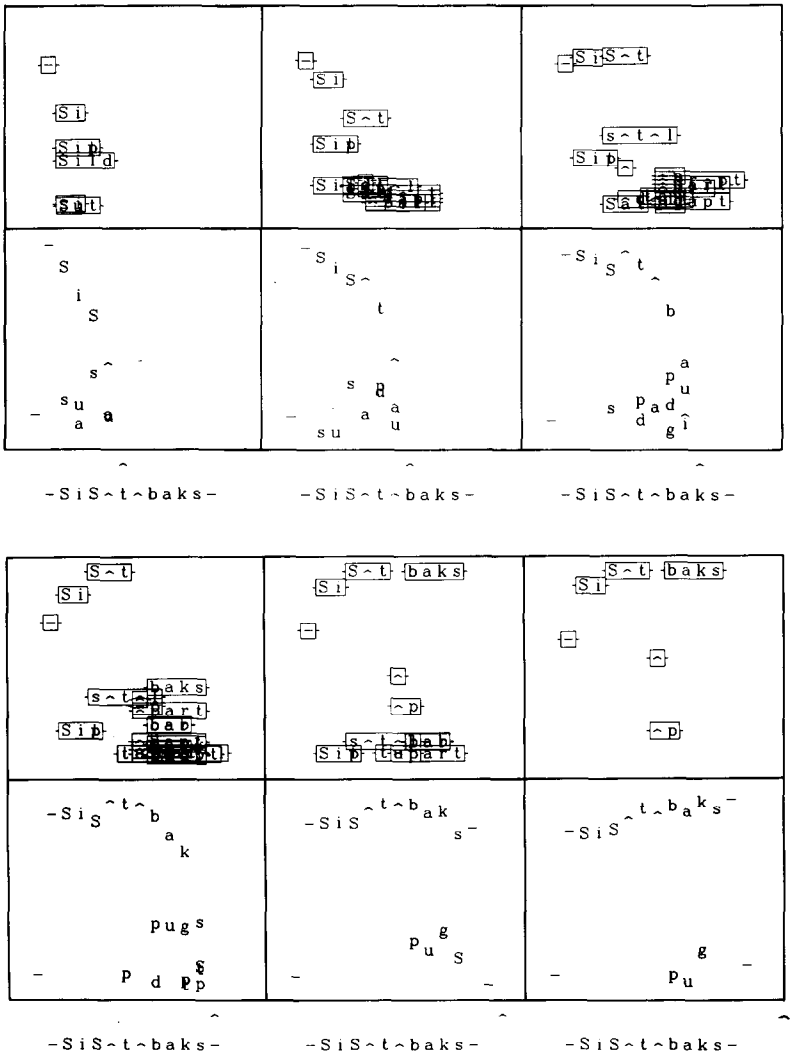

FIG. 32. The state of the Trace at several points during the processing of the stream /Sis^t^baks/ ("She shut a box").

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.

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 article, 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 which 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. These characteristics of human speech perception have been demonstrated in a very large number of studies.
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. In processing unambiguous phoneme sequences preceded by si-

lence, 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.

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

8. 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.

9. 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.

10. 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.

In addition to these characteristics observed in the present paper, our simulations with TRACE I show several further correspondences between the model and human speech perception. Most important of these is the fact 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).

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.

Some of the Reasons for the Successes of TRACE

To what does the TRACE model owe its success in simulating human speech perception? Some of TRACE's successes simply depend on its ability to make use of the information as it comes it. For example, it fails to show context effects only when a response must be made, or can be made with high accuracy, before contextual information is available.

There are several other reasons for TRACE's success. One, we think, is the use of continuous activation and competition processes in place of

discrete decisive processes such as segmentation and labeling. Activation and competition are matters of degree and protect TRACE from catastrophic commitment in marginal cases, and they provide a natural means for combining many different sources of information. Of course, this feature of the model is shared with several other models (e.g., Morton, 1969; Oden & Massaro, 1978), though only Nusbaum and Slowiaczek (1982) have previously incorporated these kinds of assumptions in a model of the time course of word recognition.

Part of the success of TRACE is specifically due to the use of competitive inhibitory interactions instead of bottom-up (or top-down) inhibition. Competition allows the model to select the best interpretation available, settling for an imperfect one when no better one is available, but overriding poor ones when a good one is at hand. These and other virtues of competitive inhibition have been noted before (e.g., Feldman & Ballard, 1982; Grossberg, 1973; Levin, 1976; Ratliff, 1965; von Bekesy, 1967) in other contexts. Their usefulness here attests to the general utility of the competitive inhibition mechanism.

The elimination of between-level inhibition from the interactive activation mechanism puts us in a very nice position with respect to one general critique of interactive-activation models. It is often said that activation models are too unconstrained and too flexible to be anything more than a language for conveniently describing information processing. We are now in a position to suggest that a restricted version of the framework is not only sufficient but superior. Interactive-activation models could exploit both excitatory and inhibitory connections both between and within levels, but in the original interactive-activation model of letter perception, only inhibitory interactions were allowed within a level. In more recent versions of the visual model (McClelland, 1985, 1986), and in TRACE, we have gone even further, allowing only excitatory connections between levels and only inhibitory connections within levels. From our experience, it appears that models which adhere to these constraints work as well as or better than members of the more general class that do not. We hasten to add that we have no proof that this is true. We have, however, no reason to feel that we could improve the performance of our model by allowing either between-level inhibitory interactions or within-level excitation.

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

We do not claim that any of these phenomena, taken individually, require the assumption of a feedback mechanism. For example, consider the phenomenon of categorical perception. We use feedback from the

phoneme to the feature level to drive feature patterns closer to the prototype of the phoneme they most strongly activate. This mechanism, coupled with the competition mechanism at the phoneme level, accounts for better discrimination between than within categories. However, we could account for categorical perception by suggesting that subjects do not have access to the acoustic level at all, but only to the results of the phoneme identification process. Similarly, lexical effects on phoneme identification can be accounted for by assuming that subjects (sometimes) read out from the word level and infer the identity of phonemes from the lexical code (Marslen-Wilson, 1980; Marslen-Wilson & Welsh, 1978; Morton, 1979). In the case of "phonotactic rule" effects, other interpretations are of course available as well. One could, for example, simply suppose that subjects use knowledge of the phonotactic constraints, perhaps captured in units standing for legal phoneme pairs, and that it is the output of such units that accounts for the influence of phonotactic regularity on phoneme identification.

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

A second reason for retaining feedback comes up when we consider the problem of learning. Although we have not discussed how learning might occur in TRACE, we have assumed that the mechanisms of speech perception are acquired through modification of connection strengths. Very roughly, in many learning schemes, connections between units are strengthened when two units tend to be activated simultaneously, at the expense of connections between units that tend not to be activated at the same time (cf. Grossberg, 1978; Rosenblatt, 1962; Rumelhart & Zipser, 1985). In such schemes, however, there is a serious problem if activation is entirely bottom-up; for in that case, once a particular unit has been "tuned" to respond to a particular pattern, it is difficult to retune it; it fires when its "expected" pattern is presented, and when it fires, its

tendency to respond to that pattern only increases. Feedback provides a way to break this vicious cycle. If higher levels insist that a particular phoneme is present, then the unit for that phoneme can become activated even if the bottom-up input would normally activate some other phoneme instead; then the learning mechanism can "retune" the detector for the phoneme so that it will need to depend less on the top-down input the next time around.

In general, the use of feedback appears to place more of the intelligence required for perception and perceptual learning into the actual perceptual mechanism itself, and to make the mechanisms which exhibit this intelligence explicit. As formulated here, these mechanisms are incredibly simple; yet they appear to buy quite a lot which often gets pushed into unspecified "decision" and "postperceptual guessing" processes (e.g., Forster, 1976).

Finally, the success of TRACE also depends upon its architecture, rather than the fundamental computational principles of activation and competition, or the decision to include feedback. By architecture, we mean the organization of the Trace structure into layers consisting of units corresponding to items occurring at particular times within the utterance. As we noted in the introduction, this architecture is one we decided upon only after several other kinds of architecture had failed.

There are three principle positive consequences of the TRACE architecture. First, it keeps straight what occurred when in the speech stream. Competition occurs only between units competing to represent the same portion of the input stream. Multiple copies of the same phoneme and word units can be active at the same time without producing confusion. Furthermore, the architecture permits the same competition mechanism that chooses among alternative word interpretations of a single-word utterance to segment longer utterances into words. No separate control structure, resetting the mechanism at the beginning of each new word, is required.

Second, the architecture permits both forward and backward interactions. Backward interactions are absolutely essential if the model is to account for the fact that the identity of a phoneme (or a word; Warren & Sherman, 1974) can be influenced by what comes after it as well as what comes before it. Some kind of record of the past is necessary to capture these kinds of influences, as well as to provide a clear picture of the sources of the more conventional effects of preceding context, and the Trace construct lays this out in a way that is both comprehensible and efficient.

Third, the Trace structure provides an explicit mechanism which instantiates the idea that there may be no distinction between the mechanisms which carry out perceptual processing and those which provide a

working memory for the results of the perceptual process. At one and the same time, the Trace is a perceptual processing system and a memory system. As a result, the model automatically accounts for the fact that coherent memory traces persist longer than incoherent ones. The coherent ones resonate through interactive (that is, bottom-up and top-down) activation, while incoherent ones fail to establish a resonance and therefore die away more rapidly.

Several of these aspects of TRACE overlap with assumptions made in other models, as mentioned in previous sections; continuity between working memory and the perceptual processing structures has been suggested by a number of other authors (e.g., Conrad, 1962), and the notion that working memory is a dynamic processing structure rather than a passive data structure has previously been advocated by Crowder (1978, 1981) and Grossberg (1978). Indeed, Grossberg has noted that resonating activation/competition processes can both enhance a perceptual representation and increase the retention of a representation; his analysis of interactive-activation processes in perception and memory captures the continuity of perception and memory as well as many other desirable properties of interactive-activation mechanisms.

Some Deficiencies of TRACE

Although TRACE has had a number of important successes, it also has a number of equally important deficiencies. A number of these deficiencies relate to simplifying assumptions of the simulation model. It is important to be clear that such deficiencies are not intrinsic to the basic structure of the model but to the simplifications we have imposed upon it to increase our ability to understand its basic properties. Certain deficiencies—such as the assumption that all phonemes are the same length, that all features are equally salient and useful and overlap an equal amount from one phoneme to another—are not present in TRACE I. Obviously a fully realistic model would take account of such differences. Other factors that should be incorporated in a more complete model include some provision for effects of word frequency, and some mechanisms for exploiting available cues to word boundaries.

Another deficiency of the model is that the decision mechanisms have not been fully enough elaborated. For example, as it stands the model does not provide a mechanism for deciding when a nonword has been presented. Nor have we specified how decision processes would actually use the information available at the word level to locate word-initial phonemes. A related problem is the lack of an explicit provision for variability in the activation and/or readout processes. Incorporating variability directly into a simulation model would greatly increase the complexity of the simulation process, but would also increase the model's

ability to capture the detailed properties of reaction time distributions and errors (Ratcliff, 1978).

So far we have considered deficiencies which we would attribute to simplifying assumptions adopted to keep TRACE as simple and transparent in its behavior as possible. However, there are some problems that are intrinsic to the basic structure of the model.

One fundamental deficiency of TRACE is that fact that it requires massive duplication 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. As we already noted, learning in activation models (e.g., Ackley, Hinton, & Sejnowski, 1985; Grossberg, 1976; Rumelhart & Zipser, 1985) usually involves the retuning of connections between units depending on their simultaneous activation. 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 returning of feature–phoneme and phoneme–word connections to take account of changes in global parameters or speaker characteristics.

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

We are currently beginning to develop a model that has these properties, based on a scheme for using a central network of units to tune the connections between the units in the Trace in the course of processing, thereby effectively programing it "on the fly." Similar ideas have already been applied to visual word recognition (McClelland, 1985, 1986). Our hope is that a new version of the model based on these ideas will preserve the positive features of TRACE I and TRACE II, while overcoming their principle deficiencies.

Some General Issues in Speech and Language Perception

There are a number of general issues in speech and language perception. Four questions in particular appear to lie close to the heart of our conception of what speech perception is all about. First, what are the basic units in speech perception? Second, what is the percept, and which aspects of the processing of spoken language should be called perceptual? Third, what is the representation of linguistic rules? Fourth, is there anything unique or special about speech perception? We conclude this article by considering each issue from the perspective we have developed through the course of our explorations of TRACE.

What is the perceptual unit? Throughout this article, we have considered three levels of processing—feature, phoneme, and word. At each level, individual processing units stand for hypotheses about the features, phonemes, and words that might be present at different points in the input stream. It is worth noting that most aspects of the model's performance are independent of the specific assumptions that we have made about the units, or even the levels. Thus, if we replaced the phoneme level with demisyllables (Fujimura & Lovins, 1978) or phoneme triples (Wickelgren, 1969), very little of the behavior of the model would change. These units can capture some of the coarticulatory influences on phoneme identity, and they would reduce some of the word-boundary ambiguities faced by the current version of the model, but neither coarticulatory influences nor word boundary ambiguities would disappear altogether (see Elman & McClelland, in press, for further discussion).

In fact, interactive activation models like TRACE can be formulated in which each perceptual object is represented, not by a single unit, but by a pattern of activation over a collection of units. For example, the phoneme units in each time slice of TRACE might be replaced by a different set of units which did not have a one-to-one correspondence to phonemes. A phoneme would be represented by a particular pattern of activation over the set of units (each representing, perhaps, to some conjunction of lower level features) rather than by a single unit in the set.

There are some computational advantages of distributed representation compared to our "one unit one concept" assumption (Hinton, Mc-

Clelland, & Rumelhart, in press), but it is very difficult to find principled ways of distinguishing between local and distributed representational schemes empirically. Indeed, in certain cases there is an exact mapping and, in general, it is possible to approximate most aspects of the behavior of a local scheme with a distributed one and vice versa (Smolensky, 1986). In light of this, our use of local as opposed to distributed representations is not perhaps as significant as it might appear at first glance. What is essential is the information that the representation captures, rather than whether it does so via distributed or local representation. The use of local representations, with each unit (at the phoneme and word levels, anyway) representing a mutually exclusive alternative makes it much easier to relate the states of the processing system to overt response categories but is not otherwise a fundamental feature of the structure of the model.

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

For one thing, TRACE blurs the distinction between perception and other aspects of cognitive processing. There is really no clear way in TRACE to say where perceptual processing ends and conceptual processes or memory begin. However, following Marr's (1982) definition of visual perception, we could say that speech perception is the process of forming representations of the stimulus—the speaker's utterance—at several levels of description. TRACE provides such a set of representations, as well as processes to construct them. On this view, then, the Trace is the percept, and interactive activation is the process of perception.

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

On the other hand, the definition seems overly liberal, for there is evidence suggesting that perceptual experience and access to the results of perceptual processing for the purposes of overt responding may not be completely unconstrained. A number of experiments, both in speech (e.g., Foss & Swinney, 1973; McNeil & Lindig, 1973) and reading (Drewnowski & Healy, 1977; Healy, 1976) suggest that under certain conditions lower levels of processing are inaccessible, or are at best accessed only

with extra time or effort. On this evidence, if perception is to form representations, and if the representations are anything like those postulated in TRACE, then perception is quite independent of the experience of the perceiver and of access to the percept. Put another way, we may choose to define the Trace as the percept, but it is not the perceptual experience. This does not seem to be a very satisfactory state of affairs.

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

How are rules represented? It is common in theories of language to assume without discussion that linguistic rules are represented *as such* in the mind of the perceiver, and that perception is guided primarily by consultation of such rules. However, there are a number of difficulties associated with this view. First, it does not explain how exceptions are handled; it would seem that for every exception, there would have to be a special rule that takes precedence over the more general formulation. Second, it does not explain aspects of rule acquisition by children learning language, particularly the fact that rules appear to be acquired, at least to a large extent, on a word by word basis; acquisition is marked by a gradual spread of the rule from one lexical item or set of lexical items to others. Third, it does not explain how rules come into existence historically; as with acquisition, it appears that rules spread gradually over the lexicon. It is difficult to reconcile several of these findings with traditional rule-based accounts of language knowledge and language processing.

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

What is special about speech? We close by raising a question that often

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

Over the last several years, a number of empirical arguments have been put forward that suggest that perhaps speech may not be so special, or at least, not unique. Cue trade-offs and contextual influences are, of course, present in many other domains (Medin & Barsalou, in press), and a large number of studies have reported categorical perception in other modalities (see Repp, 1984, for a discussion). Computational work on problems in vision have made clear that information that must be extracted from visual displays is often complexly encoded with other information (Barrow & Tenenbaum, 1978; Marr, 1982), and the lack of clear boundaries between perceptual units in vision is notorious (Ballard et al., 1983; Marr, 1982). Thus, the psychological phenomena that characterize human speech perception, and the computational problems that must be met by any mechanism of speech perception, are not, in general, unique to speech. To be sure, the particular constellation of problems that must be solved in speech perception is different than the constellation of problems faced in any other particular case, but most of the individual problems themselves do appear to have analogs in other domains.

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

REFERENCES

Ackley, D., Hinton, G., & Sejnowski, T. (1985). Boltzmann machines: Constraint satisfaction networks that learn. *Cognitive Science,* 9, 113–147.

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

Bagley, W. C. (1900). The apperception of the spoken sentence: A study in the psychology of language. *American Journal of Psychology*, **12**, 80–130.

Ballard, D. H., Hinton, G. E., & Sejnowski, T. J. (1983). Parallel visual computation. *Nature (London)*, **306**, 21–26.

Barrow, H. G., & Tenenbaum, J. M. (1978). In A. R. Hanson & E. M. Riseman (Eds.), *Computer vision systems* (pp. 3–26). New York: Academic Press.

von Bekesy, G. (1967). *Sensory inhibition*. Princeton, NJ: Princeton Univ. Press.

Bond, Z. S., & Garnes, S. (1980). Misperceptions of fluent speech. In R. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Erlbaum.

Broadbent, D. E. (1967). Word frequency effect and response bias. *Psychological Review*, **74**, 1–15.

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

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

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

Cole, R. A., & Rudnicky, A. (1983). What's new in speech perception? The research and ideas of William Chandler Bagley, 1874–1946. *Psychological Review*, **90**, 94–101.

Conrad, R. (1962). An association between memory errors and errors due to acoustic masking of speech. *Nature (London)*, **196**, 1314–1315.

Cotton, S., & Grosjean, F. (1984). The gating paradigm: A comparison of successive and individual presentation formats. *Perception & Psychophysics*, **35**, 41–48.

Crowder, R. G. (1978). Mechanisms of auditory backward masking in the stimulus suffix effect. *Psychological Review*, **85**, 502–524.

Crowder, R. G. (1981). The role of auditory memory in speech perception and discrimination. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 167–179). New York: North-Holland.

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

Drewnowski, A., & Healy, A. (1977). Detection errors on *the* and *and:* Evidence for readings units larger than the word. *Memory & Cognition*, **5**, 636–647.

Elman, J. L. (1983). Unpublished results.

Elman, J. L., & McClelland, J. L. (1984). The interactive activation model of speech perception. In Norman Lass (Ed.), *Language and speech* (pp. 337–374). New York: Academic Press.

Elman, J. L., & McClelland, J. L. (in press). Exploiting the lawful variability in the speech wave. In J. S. Perkell, & D. H. Klatt (Eds.), *Invariance and variability of speech processes*. Hillsdale, NJ: Erlbaum.

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

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

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. Walker (Eds.), *New approaches to language mechanisms*. Amsterdam: North-Holland.

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

Foss, D. J., & Gernsbacher, M. A. (1983). Cracking the dual code: Toward a unitary model of phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, **22**, 609–633.

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

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception & Psychophysics*, **36**, 359–368.

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

Fox, R. A. (1984). Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 526–540.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech*, **5**, 171–189.

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

Fujisaki, H., & Kawashima, T. (1968, August). The influence of various factors on the identification and discrimination of synthetic speech sounds. *Reports of the 6th International Congress on Acoustics*, Tokyo.

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

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

Grosjean, F. (1985). The recognition of a word after its acoustic offset: Evidence and implications. Working paper, Northeastern University, Boston.

Grosjean, F., & Gee, J. (1984). Another view of spoken word recognition. Working paper, Northeastern University, Boston.

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

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

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

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

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. I*. Cambridge, MA: Bradford Books.

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

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

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

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

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

Lane, H. L. (1965). The motor theory of speech perception: A critical review. *Psychological Review* **72,** 275–309.

Lehiste, I. (1960). An acoustic–phonetic study of internal open juncture. *Phonetica,* **5,** 1–54.

Lehiste, I. (1964). Juncture. *Proceedings of the 5th International Congress of Phonetic Sciences, Munster* (pp. 172–200). Basel/New York: S. Karger.

Levin. J. A. (1976). *Proteus: An activation framework for cognitive process models* (ISI/WP-2). Marina del Rey, CA: Information Sciences Institute.

Liberman, A. M. (1970). The grammars of speech and language. *Cognitive Psychology,* **1,** 301–323.

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

Licklider, J. C. R., & Miller, G. A. (1951). The perception of speech. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology*. New York: Wiley.

Lowerre, B. T. (1976). *The HARPY speech recognition system.* Unpublished doctoral dissertation, Carnegie–Mellon University, Pittsburgh.

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

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [s]–[š] distinction. *Perception & Psychophysics,* **28,** 213–228.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Marslen-Wilson, W. D. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature (London),* **244,** 522–523.

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

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

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition,* **8,** 1–71.

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

Massaro, D. W. (1975). *Experimental psychology and information processing.* Chicago: Rand McNally.

Massaro, D. W. (1981). Sound to representation: An information-processing analysis. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 181–193). New York: North-Holland.

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

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

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

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

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review,* **86,** 287–330.

McClelland, J. L. (1985). Putting knowledge in its place: A scheme for programming parallel processing structures on the fly. *Cognitive Science*, **9**, 113–146.

McClelland, J. L. (1986). The programmable blackboard model of reading. In J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2*. Cambridge, MA: Bradford Books.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception, Pt. I: An account of basic findings. *Psychological Review*, **88**, 375–407.

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

Medin, D. L., & Barsalou, L. W. (in press). Categorization processes and categorical perception. In S. Harnad (Ed.), *Categorical perception*. Cambridge, England: Cambridge Univ. Press.

Miller, G., Heise, G., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, **41**, 329–335.

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

Miller, J. L., Green, K., & Schermer, T. M. (1984). A distinction between the effects of sentential speaking rate and semantic congruity on word identification. *Perception & Psychophysics*, **36**, 329–337.

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

Morton, J. (1979). Word recognition. In J. Morton & J. C. Marshall (Eds.), *Psycholinguistics 2: Structures and processes* (pp. 107–156). Cambridge, MA: MIT Press.

Nakatani, L., & Dukes, K. (1977). Locus of segmental cues for word juncture. *Journal of the Acoustical Society of America*, **62**, 714–719.

Norris, D. (1982). Autonomous processes in comprehension: A reply to Marslen-Wilson and Tyler. *Cognition*, **11**, 97–101.

Nusbaum, H. C., & Slowiaczek, L. M. (1982). An activation model of auditory word recognition. *Research on speech perception, progress rep. No. 8* (pp. 289–305). Department of Psychology, Indiana University.

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

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, **13**, 253–260.

Pisoni, D. B. (1975). Auditory short-term memory and vowel perception. *Memory & Cognition*, **3**, 7–18.

Pisoni, D., & Lazarus, J. H. (1974). Categorical and non-categorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, **55**, 328–333.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, **15**, 285–290.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, **85**, 59–108.

Ratliff, F. (1965). *Mach bands: Quantitative studies on neural networks in the retina*. San Francisco: Holden Day.

Reddy, D. R. (1976). Speech recognition by machine: A review. *Proceedings of the IEEE*, **64**, 501–531.

Reddy, D. R., Erman, L. D., Fennell, R. D., & Neely, R. B. (1973). The Hearsay speech

understanding system: An example of the recognition process. *Proceedings of the International Conference on Artificial Intelligence* (pp. 185–194). Stanford, CA.

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In J. Lass (Ed.), *Speech and Language* (Vol. 10). New York: Academic Press.

Repp, B. H., & Liberman, A. M. (1984). Phonetic categories are flexible. *Haskins Laboratories Status Report on Speech Research*, SR-77/78, 31–53.

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

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

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception, Pt. II: The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, **89**, 60–84.

Rumelhart, D. E., & Zipser, D. (1985). Competitive learning. *Cognitive Science*, **9**, 75–112.

Salasoo, A., & Pisoni, D. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory and Language*, **24**, 210–231.

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

Smolensky, P. (1986). Neural and conceptual interpretation of PDP models. In J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. II*. Cambridge, MA: Bradford Books.

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

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

Summerfield, Q., & Haggard, M. (1977). On the dissociation of spatial and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America*, **62**, 435–448.

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

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

Tyler, L. K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics*, **34**, 409–420.

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

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

Wolf, J. J., & Woods, W. A. (1978). The HWIM speech understanding system. In W. A. Lea (Ed.), *Trends in speech recognition*. Englewood Cliffs, NJ: Prentice-Hall.

(Accepted July 25, 1985)