COGNITIVE PSYCHOLOGY 18, 1–86 (1986) 0010-0285/86 $7.50
Copyright © 1986 by Academic Press, Inc.
All rights of reproduction in any form reserved.

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 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. © 1986 Academic Press, Inc.

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.

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

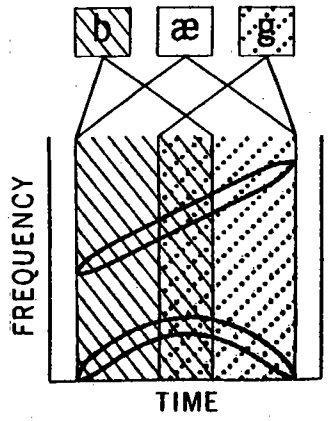

FIG. 1. A schematic spectrogram for the syllable "bag," indicating the overlap of the information specifying the different phonemes. Reprinted with permission from Liberman (1970).

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,

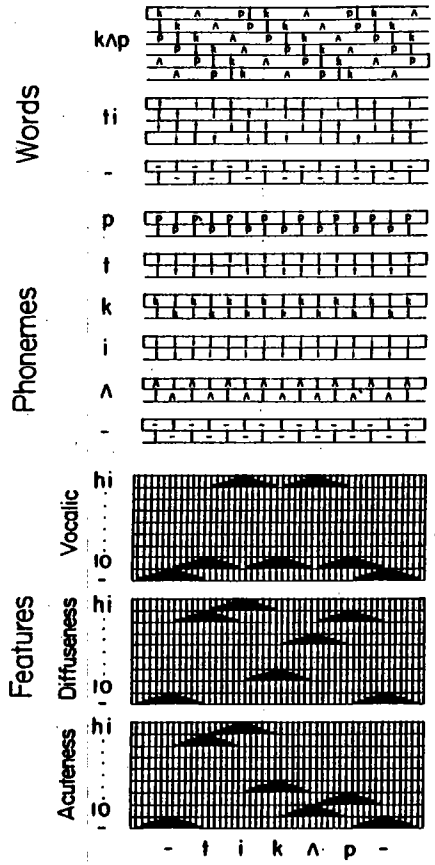

FIG. 2. A subset of the units in TRACE II. Each rectangle represents a different unit. The labels indicate the item for which the unit stands, and the horizontal edges of the rectangle indicate the portion of the Trace spanned by each unit. The input feature specifications for the phrase "tea cup," preceded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

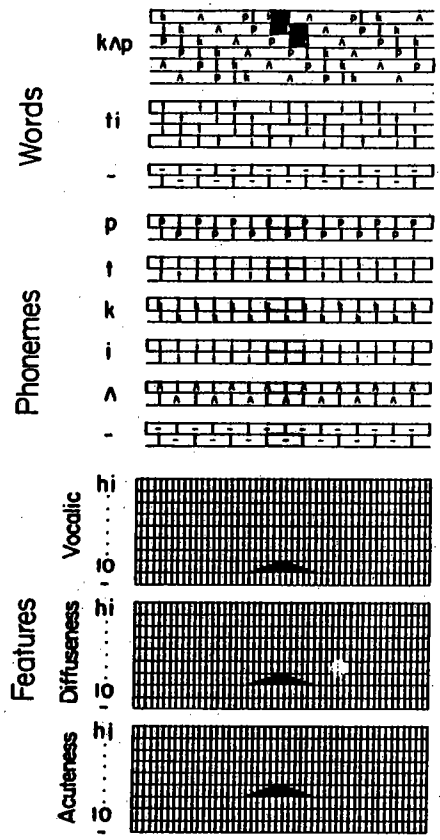

FIG. 3. The connections of the unit for the phoneme /k/, centered over Time Slice 24. The rectangle for this unit is highlighted with a bold outline. The /k/ unit has mutually excitatory connections to all the word- and feature-level units colored either partly or wholly in black. The more coloring on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

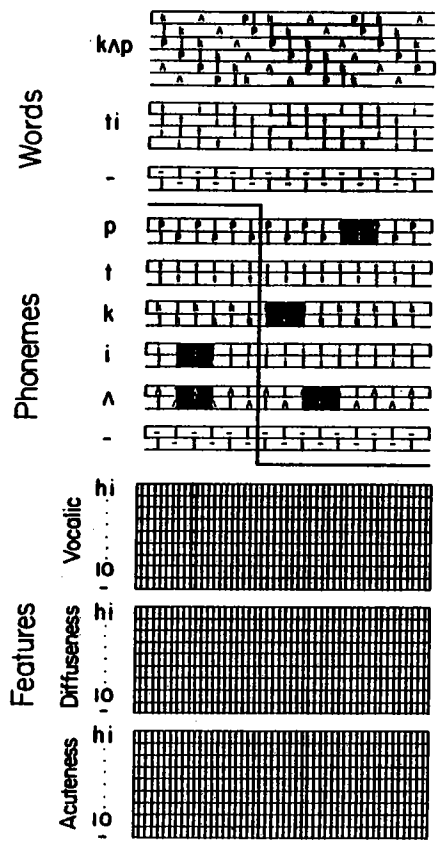

FIG. 4. The connections of the highlighted unit for the high value on the Vocalic feature dimension in Time Slice 9 and for the highlighted unit for the word /kˆp/ starting in Slice 24. Excitatory connections are represented in black, inhibitory connections in grey, as in Fig. 3.

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	—						
r									—	.80	.29		.20	.37
l									.80	—	.32		.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 /ˆ/ 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 cycles 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.