CHAPTER 15

Interactive Processes in Speech Perception: The TRACE Model

J. L. McCLELLAND and J. L. ELMAN

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

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

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

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

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

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

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

SOME IMPORTANT FACTS ABOUT SPEECH

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

Temporal Nature of the Speech Stimulus

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

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

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

Left and Right Context Effects

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

Lack of Boundaries and Temporal Overlap

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

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

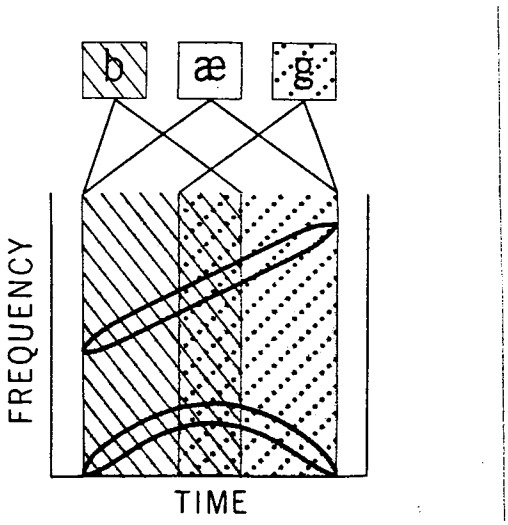

FIGURE 1. A schematic spectrogram for the syllable *bag*, indicating the overlap of the information specifying the different phonemes. (From "The Grammars of Speech and Language" by A. M. Liberman, 1970, *Cognitive Psychology*, *1*, p. 309. Copyright 1970 by Academic Press, Inc. Reprinted by permission.)

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

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

Context Sensitivity of Cues

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

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

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

Noise and Indeterminacy in the Speech Signal

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

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

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

The Importance of the Right Architecture

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

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

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

THE TRACE MODEL

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

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

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

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

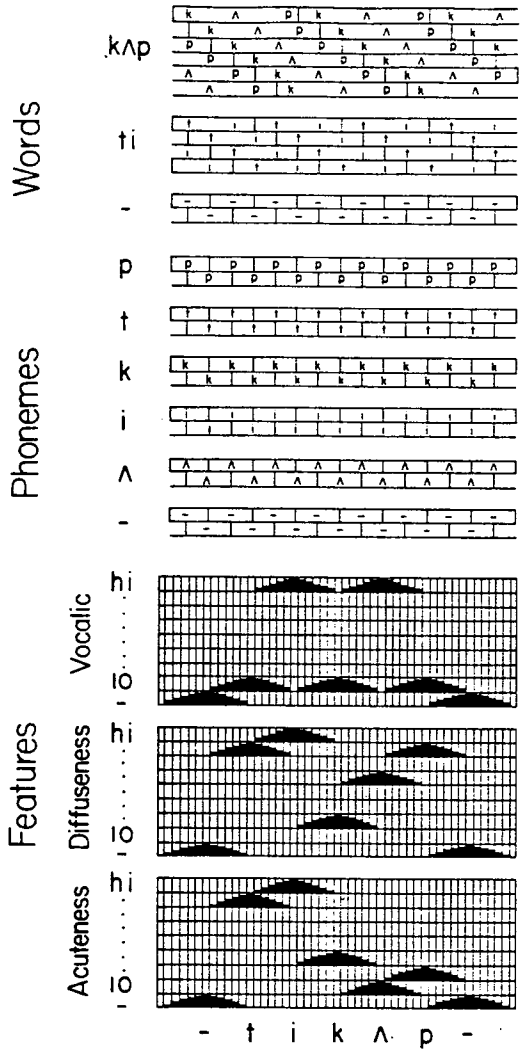

FIGURE 2. A subset of the units in TRACE II. Each rectangle represents a different unit. The labels indicate the item for which the unit stands, and the horizontal edges of the rectangle indicate the portion of the Trace spanned by each unit. The input feature specifications for the phrase *tea cup*,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

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

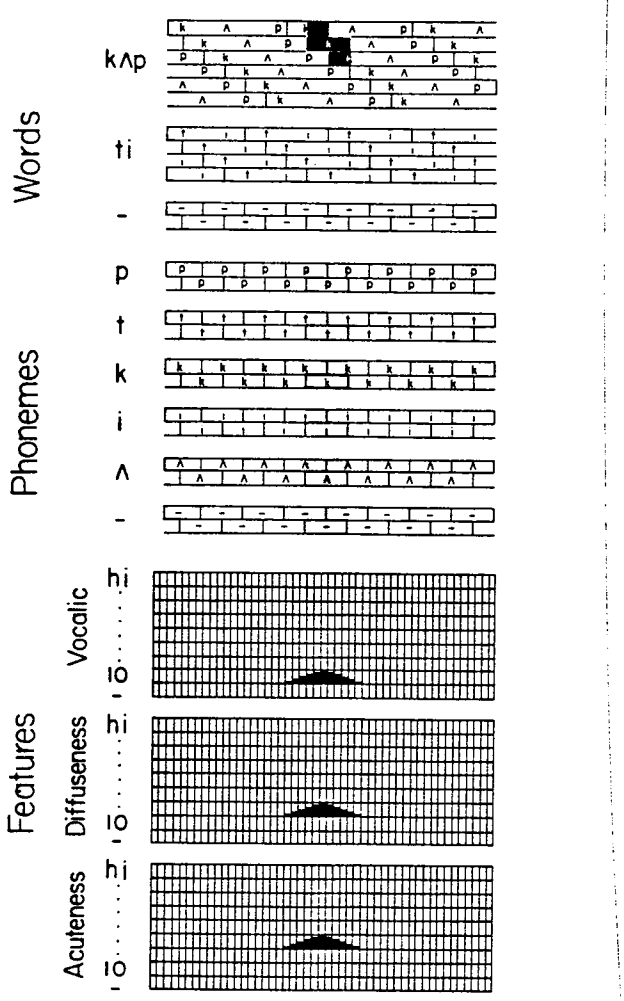

FIGURE 3. The connections of the unit for the phoneme /k/, centered over Time-Slice 24. The rectangle for this unit is highlighted with a bold outline. The /k/ unit has mutually excitatory connections to all the word- and feature-level units colored either partly or wholly in black. The more coloring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

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

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

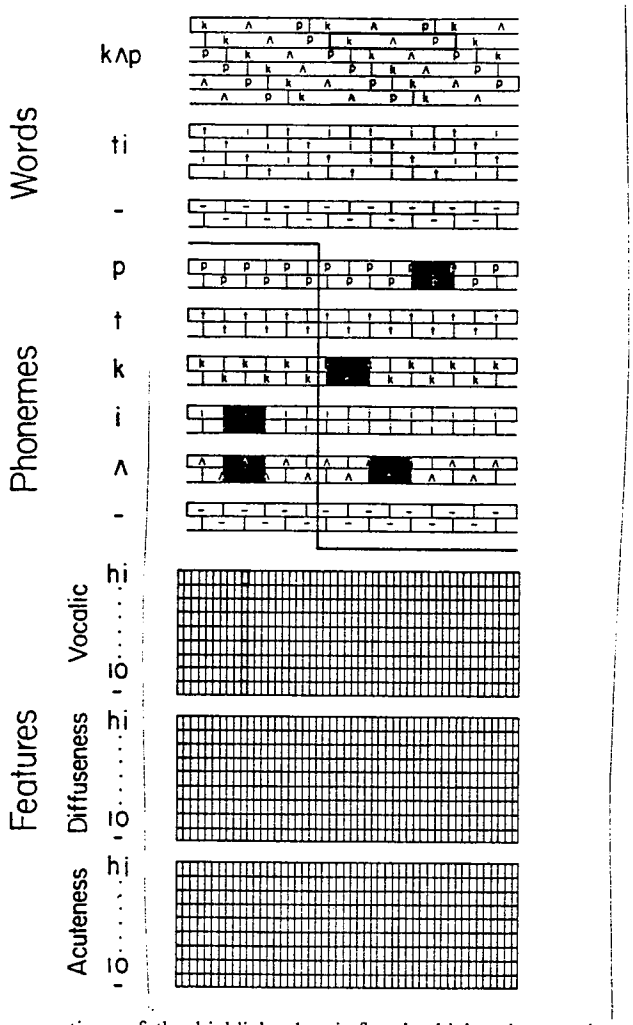

FIGURE 4. The connections of the highlighted unit for the high value on the vocalic feature dimension in Time-Slice 9 and for the highlighted unit for the word /kˆp/ starting in Slice 24. Excitatory connections are represented in black, inhibitory connections in gray, as in Figure 3.

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

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

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

Context-Sensitive Tuning of Phoneme Units

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

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

TRACE I and TRACE II

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

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

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

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

Implementation Details

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

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

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

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

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

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

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

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

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

TABLE 1

PHONEME FEATURE VALUES USED IN TRACE II

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

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

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

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

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

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

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

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

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

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

It is important to remember that the input is applied, one slice at a time, proceeding from left to right as though it were an ongoing stream of speech "writing on" the successive time slices of the Trace. The interactive activation process is occurring throughout the Trace on each time slice, even though the external input is only coming in to the feature units one slice at a time. Processing interactions can continue even after the left to right sweep through the input reaches the end of the Trace. Once this happens, there are simply no new input specifications applied to the Trace; the continuing interactions are based on what has already been presented. This interaction process is assumed to continue indefinitely, though for practical purposes it is always terminated after some predetermined number of time cycles has elapsed.

Activations and overt responses. Activations of units in the Trace rise and fall as the input sweeps across the feature level. At any time, a decision can be made based on the pattern of activation as it stands at that moment. The decision mechanism can, we assume, be directed to consider the set of units located within a small window of adjacent slices within any level. The units in this set then constitute the set of response alternatives, designated by the identity of the item for which the unit stands (note that with several adjacent slices included in the set, several units in the alternative set may correspond to the same overt response). Word-identification responses are assumed to be based on readout from the word level, and phoneme-identification responses are assumed to be based on readout from the phoneme level.

As far as phoneme identification is concerned, then, a homogeneous mechanism is assumed to be used with both word and nonword stimuli. The decision mechanism can be asked to make a response either (a) at a critical time during processing, or (b) when a unit in the alternative set reaches a critical strength relative to the activation of other alternative units. Once a decision has been made to make a response, one of the alternatives is chosen from the members of the set. The probability of choosing a particular alternative i is then given by the Luce (1959) choice rule:

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

where j indexes the members of the alternative set, and $S_i = e^{ka_i}$. The exponential transformation ensures that all activations are positive and gives great weight to stronger activations; the Luce rule ensures that the sum of all of the response probabilities adds up to 1.0. Substantially the same assumptions were used by McClelland and Rumelhart (1981).

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

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

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

TABLE 2

PARAMETERS OF TRACE II

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

*Per 3 time slices of overlap.

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

FACTORS INFLUENCING PHONEME IDENTIFICATION

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

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

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

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

Lexical Effects

You can tell a phoneme by the company that it keeps.[1] In this section, we describe a simple simulation of the basic lexical effect on

[1] This title is adapted from the title of a talk by David E. Rumelhart on related phenomena in letter perception. These findings are described in Rumelhart and McClelland (1982).

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

For the first simulation, the input to the model consisted of a feature specification which activated /b/ and /p/ equally, followed by (and partially overlapping with) the feature specifications for /l/, then /ʌ/, then /g/. Figure 5 shows phoneme- and word-level activations at several points in the unfolding of this input specification. Each panel of the figure represents a different point in time during the presentation and concomitant processing of the input. The upper portion of each panel is used to display activations at the word level; the lower panel is used for activations at the phoneme level. Each unit is represented by a rectangle labeled with the identity of the item the unit stands for. The horizontal extension of the rectangle indicates the portion of the input spanned by the unit. The vertical position of the rectangle indicates the degree of activation of the unit. In this and subsequent figures, activations of the phoneme units located between the peaks of the input specifications of the phonemes (at Slices 3, 9, 15, etc.) have been deleted from the display for clarity. The input itself is indicated below each panel, with the successive phonemes positioned at the temporal positions of the centers of their input specifications. The "^" along the x-axis represents the point in the presentation of the input stream at which the snapshot was taken.

The figure illustrates the gradual build-up of activation of the two interpretations of the first phoneme, followed by gradual build-ups in activation for subsequent phonemes. As these processes unfold, they begin to produce word-level activations. It is difficult to resolve any word-level activations in the first few frames, however, since in these frames, the information at the phoneme level simply has not evolved to the point where it provides enough constraint to select any one particular word. It is only after the /g/ has come in that the model has information telling it whether the input is closer to *plug*, *plus*, *blush*, or *blood* (TRACE's lexicon contains no other words beginning with /plʌ/ or /blʌ/). After that point, as illustrated in the fourth panel, *plug* wins the competition at the word level, and through feedback support to /p/, causes /p/ to dominate /b/ at the phoneme level. The model, then, provides an explicit account for the way in which lexical information can influence phoneme identification.

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

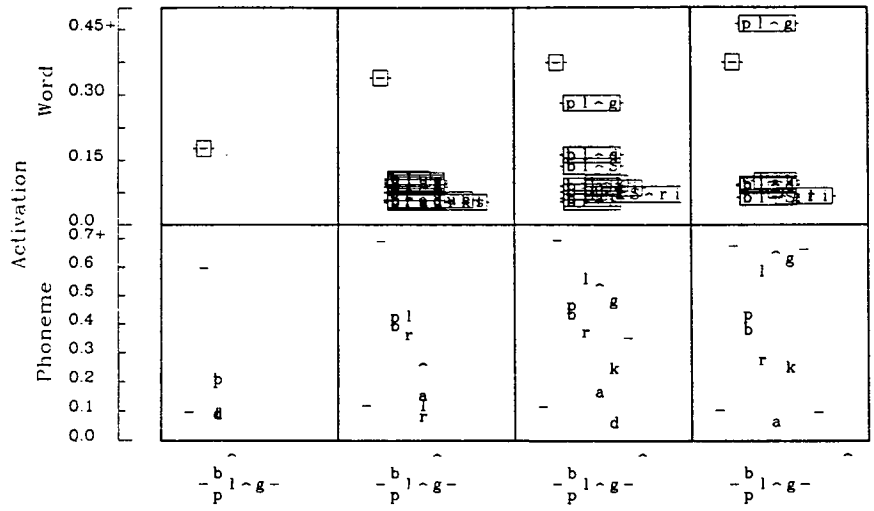

FIGURE 5. Phoneme- and word-level activations at several points in the unfolding of a segment ambiguous between /b/ and /p/, followed by /l/, /ˆ/, and /g/. See text for a full explanation.

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

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