

The TRACE Model of Speech Perception

JAMES L. MCCLELLAND

Carnegie-Mellon University

AND

JEFFREY L. ELMAN

University of California, San Diego

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

The work reported here was supported in part by a contract from the Office of Naval Research (N-00014-82-C-0374), in part by a grant from the National Science Foundation (HNS-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 programming and graphics. Send requests for reprints to James L. McClelland, Department of Psychology, Carnegie-Mellon University, Schenley Park, Pittsburgh, PA 15213.

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.

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 Rudnicki (1983) have recently noted, these basic facts were described in early experiments by Bagley (1900) over 80 years ago. Cole and Rudnicki point out that recent work (which we consider in detail below) has added clarity and detail to these basic findings but has not led 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 artificial 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.

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 (Lieberman, 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 demissyllables (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

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 (Lieberman, 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 Dukas (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

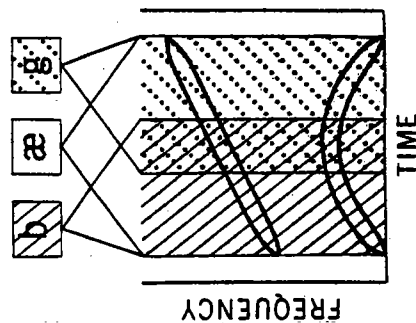


FIG. 1. A schematic spectrogram for the syllable "bag," indicating the overlap of the information specifying the different phonemes. Reprinted with permission from Liberman (1970).

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 Eelman & 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" (/tik p/) 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,

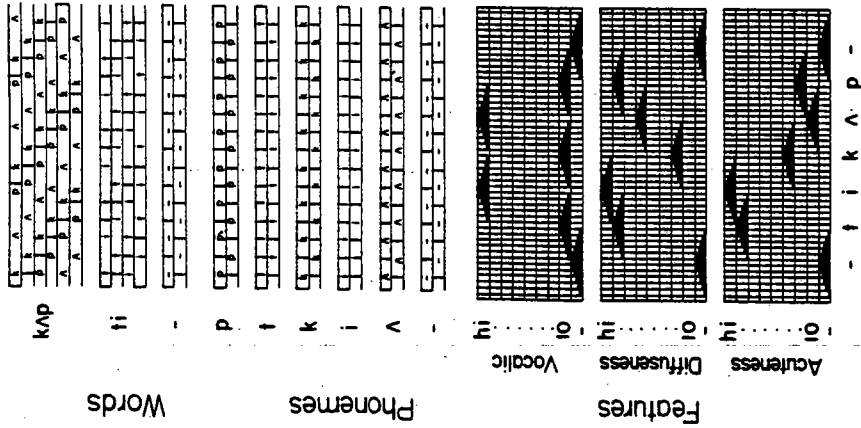


FIG. 2. A subset of the units in TRACE II. Each rectangle represents a different unit. The labels indicate the item for which the unit stands, and the horizontal edges of the rectangle indicate the portion of the Trace spanned by each unit. The input feature specifications for the phrase "tea cup," preceded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

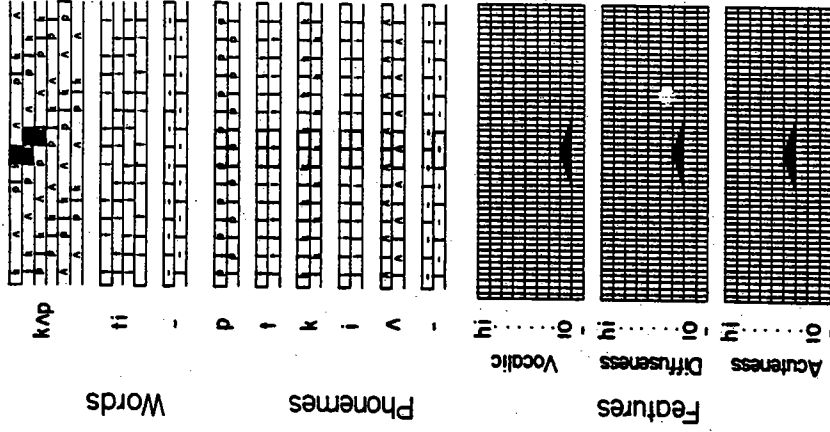


FIG. 3. The connections of the unit for the phoneme /k/. centered over Time Slice 24. The rectangle for this unit is highlighted with a bold outline. The /k/ unit has mutually excitatory connections to all the word- and feature-level units colored either partly or wholly in black. The more coloring on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

