In J.S. Perkell and D.H. Klatt "Invariance and
Variability in Speech Processes", 1986, LEA/Hillsdale,
NJ

17 Exploiting Lawful Variability In the Speech Wave

effrey L. Elman and James L. McClelland
epartments of Linguistics and Psychology
niversity of California, San Diego

INTRODUCTION

recurring theme in speech perception is the enormous variability one en-
ounters in the acoustic signal. Yet listeners appear able to move beyond
is surface variability and retrieve a level of representation which consists
a smaller number of more abstract and perceptually constant elements;
oreover, this is done with apparent ease.

Of course, the variability is far from random, and this is key. Many of
e alterations in the acoustic form of speech sounds arise out of contextual
fects which are predictable, if complex. The problem is not that the sig-
al varies randomly or contains meaningless noise. It is rather that there
e many factors which need to be taken into account in the processing of
y given sound. A tremendous amount of information is conveyed simul-
neously. In our view, the phenomenon which has been described as the
ack-of-invariance problem" is not a problem at all for human listeners. It is
ecisely the variability in the signal which permits listeners to understand
eech in a variety of contexts, and spoken by a variety of speakers.[1] Instead
searching for invariance in the signal, we think it makes more sense to try
understand how it is that listeners deal with the variability which is there.

What is needed is a processing framework which lends itself to the analysis
multiple overlays of information. Machines (serial processors) are not particularly effective at integrating information from many sources, at least not in real-time. On the other hand, parallel processing systems—in particular interactive-activation models—are ideally suited to solving this problem.

[1] We are particularly concerned in this paper with acoustic variability which is predictably
nditioned by phonetic context and other linguistic factors. Cross-speaker and within-
eaker token-to-token variability, while important phenomena in speech perception, are
t directly addressed here.

In this paper we will describe our efforts at applying an interactive-activation approach to the task of performing the mapping from the digitized speech wave to a phonetic level of representation. In the process, we demonstrate how the lawful variability in the speech wave can be treated as a rich source of information, rather than as a regrettable source of noise.

THE INTERACTIVE ACTIVATION FRAMEWORK

The interactive activation framework was laid out by McClelland and Rumelhart (1981). Their paper gives a more detailed description of the basic assumptions and a discussion of the predecessors of the interactive activation approach.

Nodes

Our system is based upon a large number of simple processing elements we call "nodes." Nodes may represent different concepts (e.g., features, phonemes, or words) but are alike in being computationally very simple. Each node has an activation value, which reflects the state of the evidence that the unit the node stands for is present in the input. Each node also has a resting level, toward which its activation tends to decay; and a threshold. Nodes with activation levels below threshold do not influence other nodes. However, once the activation of a node exceeds threshold, it begins to influence the nodes to which it is connected by signaling its activation value to other nodes to which it is connected.

Connections

Connections between nodes can be of two kinds: excitatory or inhibitory. In addition, they may have different strengths associated with them. Excitatory connections have positive strengths, or weights, while inhibitory connections have negative weights. The sign determines whether the connection is excitatory or inhibitory; the absolute value of the weight determines the magnitude of the excitatory or inhibitory effect. The signal from one node to another is simply the product of the output of the sending node times the strength or weight of the connection between the nodes.

The Interactive Activation Process

The interactive activation process is the process of updating the activations of nodes based on the signals they receive from other nodes. This is an iterative process, with each node updating its activation value based on the signals it gets from the other nodes on each iteration or cycle. Each cycle consists of two phases. In the first, a net input to each node is determined, by adding up all the separate signals arising from other nodes. Then, the activations of all of the nodes are updated. The new activation of each node is just the old activation, plus the effect of the net input from other nodes, minus a decay term. If the net input to a node is positive (or excitatory), it will tend to drive the activation value of the node in a positive direction. On the other hand, if the net input is inhibitory, it will tend to drive the activation of the node in a negative direction. However, nodes have restricted dynamic ranges; for each node there is a maximum (arbitrarily set at +1.0) and a minimum (usually set at -0.2). If the net input is excitatory, it tends to drive the activation toward the maximum; if negative, toward the minimum. However, the magnitude of the effect is modulated by the distance left to go, so that the activation value always stays in bounds. The decay term tends to restore the activation to the resting level, but does so gradually over many cycles of interactive activation.

THE TRACE MODEL

Overall Structure of TRACE

The TRACE model is an interactive activation model, in that it consists of a set of nodes which interact with each other as described above. In addition to these basic assumptions, the TRACE model makes several key assumptions about the way in which the units are organized and their connections disciplined.

Levels. Nodes in the model are organized into levels. In the TRACE model, there is a feature level, consisting of detectors for acoustic/phonetic features of the speech wave; there is a phonetic level, consisting of detectors for phonemes; and there is a word level consisting of detectors for words.

Time slices. Time is an important variable in speech perception. It would appear to be a trivial observation that not only do we hear sounds, we also know when we heard them; it appears, then, that we keep a record of what units have occurred at each point in time in the input, at least for a short period of time after the input has arrived.

The TRACE. Our view is that the structures in which this perceptual memory resides and those in which current perception is carried out are one and the same. Time in the TRACE model is realized as a series of "frames" or "slices," each of which contains an identical complete set of nodes. There is one slice every 5 msec. The speech input is directed to successive time slices, one after another. Together, these slices make an active processing structure called the TRACE, in which the percept unfolds, and which serves as the working memory representation of the input.

Within each level of the TRACE, then, complete sets of detectors are reduplicated for successive time intervals in a speech stream. Thus the feature level consists of multiple complete sets of feature detectors, one for each 5 msec time-slice of the speech wave.[2] Similarly, the phoneme level consists of a complete set of phoneme detectors for each slice, and the word level consists of a complete set of word detectors for each slice. In the phoneme case, each detector stands for the hypothesis that the phoneme in question is centered below the time slice in question; in the word case, each detector stands (roughly) for the hypothesis that the first phoneme of the word is centered below the time slice in question.

Connections. The levels serve to discipline the interactions between the units. Between-level connections are exclusively excitatory, and are bidirectional. Quite simply, there is a mutually excitatory link between mutually consistent nodes on adjacent levels. Thus, feature nodes excite nodes for phonemes in which they occur in corresponding time slices, and phoneme nodes excite nodes for the features they contain. Similarly, phoneme nodes excite nodes for words in which they occur, and word nodes excite nodes for the phonemes they contain. Words are of course spread out over several phonemes, so for example the node for the word "cat" starting in a given slice will have mutually excitatory connections to [k] in the same and neighboring slices, [æ] in the next few slices, and [t] in the next few slices after that. Within-level connections are exclusively inhibitory, and are again bidirectional. Each node, when active, tends to inhibit all those nodes on the same level which are mutually inconsistent with it. Thus, for example, the nodes for different phonemes in the same time slice are mutually inhibitory. Word nodes are mutually inhibitory in proportion to the extent of their overlap in time.

Exploiting lawful variability

The values of the weights associated with the connections between the feature nodes and the phoneme nodes determine the extent to which a particular pattern of feature values will activate a particular phoneme. Context (adjecent) phonemes alter the feature patterns associated with a particular phoneme

[2] Although features are computed every 5 msec, the computation may involve reference to input from other frames. The feature ABRUPT, for instance, is calculated as the rate of change of power across a window 15 msec wide.

in a lawful way. We can exploit this regularity by allowing the detectors for these context phonemes to alter the weights appropriately for phonemes in neighboring slices, so that when the contextually-appropriate feature pattern occurs, the right phoneme will be activated. We will consider how this is done in more detail below.

A Detailed Look at the Feature and Phoneme Levels

Thus far we have described the TRACE model very generally, to give the reader a sense of its basic structure. Since in this paper we will be applying the model to the problem of exploiting the lawful variability in the speech wave, we will be focusing our attention on the feature and phoneme levels. The following section describes the input to the feature level, and is followed by two more which describe the feature and phoneme levels in more detail. Following these sections, we describe some simulations which demonstrate the ability of the model to exploit the variability in the speech wave, using the weight-modulation scheme described above.

Input to the Feature Level. The input to the feature nodes is simply the output of a speech preprocessor. In the preprocessor, digitized speech is preprocessed and converted to a set of values along each of several feature continua, at successive 5 msec. slices. Thus, the input to the model is simply a set of real-valued parameters, one for each feature continuum, for each 5 msec slice of speech.

Calculation of the feature parameter values input to the feature detectors is carried out through standard numerical techniques. We avoid incorporating this stage of processing into the model simply to speed up the simulation and not because we do not believe it can be done more profitably in an interactive activation framework. For the present we focus attention on the next stage of processing, in which feature-node activation values must be mapped to phonemes, to show how the tuning of these mappings on the basis of context can help the perceptual mechanism exploit lawful variability.

The output of the preprocessor is fed to the feature level nodes one time slice at a time, simulating the temporal flow of real speech. The successive sets of feature parameters are directed to successive slices at the feature level, so that each slice captures the parameters of a single 5 msec of speech.

Feature level. At the feature level, each feature continuum is processed by a group of eight nodes, each of which responds maximally to a different value along that feature continuum. Thus, each node may be called a feature-value detector. One such detector might respond best to very high values of a feature, while another would have a response function favoring slightly lower values. The response function for each of the eight nodes is determined by the statistical properties of our data base. We first determine the frequency of occurrence for each value along a feature continuum. This distribution is used to calculate "octiles," and then the midpoints of the octiles are used as the response peaks for the eight feature value detectors. Each detector responds best to its peak value, and the response functions fall off linearly on both sides, reaching half their maximum at the peak of the neighboring detector on either side. Thus the detectors respond to overlapping ranges of values along the continuum.

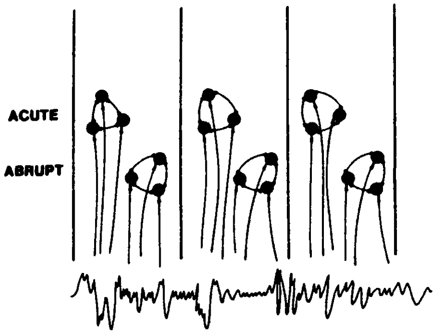

FIG. 17.1. Three time slices at the feature level. Input comes from the speech signal and excites the feature detector nodes. (Only two of 16 detector sets are shown; in reality each set has eight rather than three nodes.)

The eight nodes have two types of "outgoing" connections. There are inhibitory connections between the nodes, so that the node which responds to high values on the continuum will inhibit the nodes which detect lower values. The eight nodes also have excitatory connections to phonemes nodes.

In Figure 17.1 we illustrate how the TRACE model appears at the feature level. Three time slices are shown here. The left-most slice is the earliest in time. As time proceeds, successive slices are activated by the portion of the speech input which corresponds to that point in time. Note that the figure is simplified: only two feature banks are shown. Also, each bank is shown as having only three detectors rather than eight.

Some discussion is in order here about our choice of feature continua. Several factors governed their selection; in all cases, we assumed that feature values were continuously scaled rather than discretely valued (as in binary features). First, we needed features which have clear acoustic correlates. Note that this does not mean that the features have invariant realizations across different contexts, simply that there be some algorithm for computing the feature value at every instant in time. Second, insofar as was possible, we wished to find features which had a linguistic motivation. We used several features which have been proposed by Jakobson, Fant, and Halle (1952); these features have proven useful in describing a number of common linguistic phenomena. Third, because it is our belief that much of the variability in the

TABLE 17.1

Distinctive Features in Present Use

power	log rms power
pitch	fundamental frequency of source
alpha	total LPC error; measures predictability of signal; distinguishes speech and silence
edr	error dynamic range; voiced from voiceless
abrupt	change in rms power over time; stops/vowel transitions
diffuse	second moment of LPC spectrum
acute	gross concentration of energy in spectrum
consonantal	smoothed euclidean spectral difference; detects stop bursts and consonantal transitions, as opposed to slower changing vowel formants
spectrum[3]	total energy in each section of the spectrum divided equally in three on a log scale
formant[3]	the values of the first three formants
loci[2]	the hypothetical formant onsets, back-extrapolated from the actual formant onsets (at voicing) to the stop release; provides a context-sensitive measure of the place of articulation of consonants

speech signal can be exploited to provide additional information, we looked for features which seemed especially susceptible to variation which could be predicted on the basis of context.

Our current set of continua is described in Table 17.1. There is a bank of eight detectors for each continuum. We thus have a set of eight detectors which respond to different values of the feature ABRUPT, another eight which respond to ACUTE, and so on. We make no claim that this is the "right" set of features; indeed, we are quite sure the human perceptual system uses a much richer set. The set we have chosen does, however, allow us to begin to test or central theoretical claims about how variability may be exploited with real speech input.

Phoneme Level. The 128 feature nodes in each time slice connect to nodes at the next higher (conceptual) level; these latter nodes represent phonemes.

There has been considerable controversy about the psychological and linguistic reality of phonemes and a variety of other units have been proposed, including transemes, context-sensitive allophones, syllables, and demisyllables. We do not wish to take a strong position on this issue. We do note that many of the criticisms of the phoneme stem from its lack of invariance across different environments, and that some of the alternatives are claimed to "solve" this problem to some degree. In reality, many of these solutions are not solutions at all; they solve the problem of invariance either (1) by making the units sufficiently big that they incorporate much of the context into them; or (2) by making the units extremely small or highly context-specific. The problem with the first solution is that one can never truly "freeze in" enough of the context to guarantee an invariant form for the unit. Both solutions make it difficult to capture the intuition shared by many speakers that the acoustically different bilabial stops in "ball" and "crib" are the same sound. In any event, we feel that the approach we have taken in the TRACE model is to some extent independent from specific representational issues, and that at the very least, we can demonstrate that the lack of invariance is not, in and of itself, a reason to reject a phonemic representation, since the solution we propose to the invariance problem shows how varying input can be mapped in a contextually appropriate way onto the phoneme level.

There are several important differences between the feature and phoneme nodes. Feature values tend to change more rapidly than phoneme values and so they must be more labile. This is accomplished in two ways. First, whereas there are feature nodes for every 5 msec time slice (this seems a reasonable time grain) phoneme level time slices are spaced at 15 msec intervals. Thus phoneme-level slices span three feature level time-slice windows. Second, the receptive fields of phoneme nodes span several slices at the feature level, so that phonemes in successive slices have overlapping "windows" on the feature level. Connections between feature nodes and two phoneme nodes are shown in Figure 17.2. Stronger connections are shown as solid lines and weaker connections with dotted lines. (Note also the convention that excitatory links end in arrows and inhibitory connections end with filled circles.)

The exact nature of the feature-to-phoneme node connections is complicated by the fact that different phonemes have different durations. There are differences in the intrinsic duration of different phonemes. Thus, phoneme nodes must collect input over varying extents of feature nodes. Stop nodes, for example, might have "windows" six feature traces wide whereas vowel nodes might have windows 12 traces wide (or wider). Or the duration of

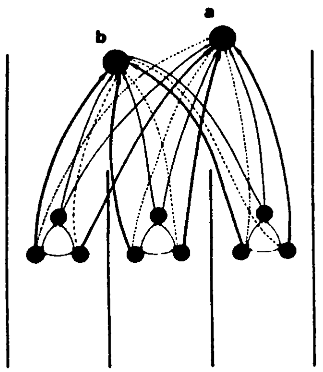

FIG. 17.2. Connections between feature detector nodes and phoneme nodes. Each detector has a connection to every phoneme, but the strength of the connection varies as is appropriate.

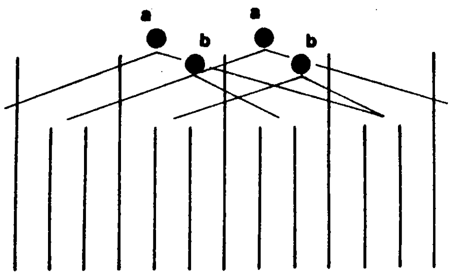

FIG. 17.3. Phoneme nodes have receptive fields which span several slices at the feature level, so that phonemes in successive slices have overlapping "windows" on the feature level.

stops may differ significantly from one another as a function of factors such as their place of articulation. In Figure 17.3 we see sample windows over which several phoneme nodes collect input.

The windows are represented by triangular functions, such that input from slices further in time from the phoneme node's "home slice" has a weaker effect than input from time slices right under the phoneme node. The area under the triangles of different widths is normalized, so that the detectors for phoneme nodes of different width are equally excitable. Without this normalization, longer phonemes would receive more excitation than shorter ones, artificially biasing the model to hear longer phonemes. Obviously this explanation only hints at a general solution to the problem of duration. One problem is that some features of the same phoneme spread further than others. We have adopted the present scheme primarily for simplicity.

Mapping from the Feature Level to the Phoneme Level. Previous versions of interactive activation models have assumed that the connections between nodes are fixed. For example, McClelland and Rumelhart (1981) in the interactive activation model of reading, had fixed excitatory connections of fixed strength between feature and letter nodes, and between letter and word nodes. However, there is a major problem with using fixed connection strengths in mapping of feature level information to the phoneme level; it is in fact the major topic of this volume. The problem is the lack of an invariant relationship between the feature values which are obtained for a given phoneme across different contexts. We have already remarked that the major problem which the lack of invariance presents is not that the variation is random or represents noise. In most cases, the causes for the variation are well-understood. Most importantly, they can often be predicted. Because of their lawful nature, it should be possible to use the context to adjust the mapping from the feature to the phoneme level.

At first this may seem like an impossibility. How can we use what we know about one phoneme to help us identify another, when each depends on the other? The answer demonstrates one of the main strengths of the interactive activation approach. These models do not have to wait until sufficient evidence has accumulated to identify a phoneme unambiguously before that phoneme can alter the processing of the phonemes around it. Interactive activation allows for on-going retuning, such that the contextual effects are immediate and direct, and grow in strength as evidence for the identity of the phonemes in the context is accumulated. An example will show how this might work.

In the model, the detector nodes for the values of the ACUTE feature project to—among other phonemes—nodes for the stop consonants [b], [d], and [g]. The strength of the connections between any given stop node and the eight ACUTE nodes depends on the observed characteristics of that stop. The alveolar stop, for instance, has stronger connections to the nodes which respond to the high end of the ACUTE feature continuum (because its spectrum is typically tilted toward the high end of the spectrum).

However, there are circumstances in which the spectrum of other stops may be shifted toward higher ACUTE values; contiguity to the vowel [i] is one such situation. In Figure 17.4 we can compare the values of ACUTEness that were detected during the stop closure of the bilabial stops [b] when it preceded different vowels. The upper trace in this figure shows how ACUTEness varies with time. The arrows indicate the moment in time when each bilabial stop is released. (For convenience in locating the stop, spectrograms are included at the bottom.) Thus the degree to which detectors for high-ACUTE should excite a given stop should depend not only on the input present during the stop but also the context. Similar sensitivities to context are also found for other feature continua, including the two LOCUS continua. Based on

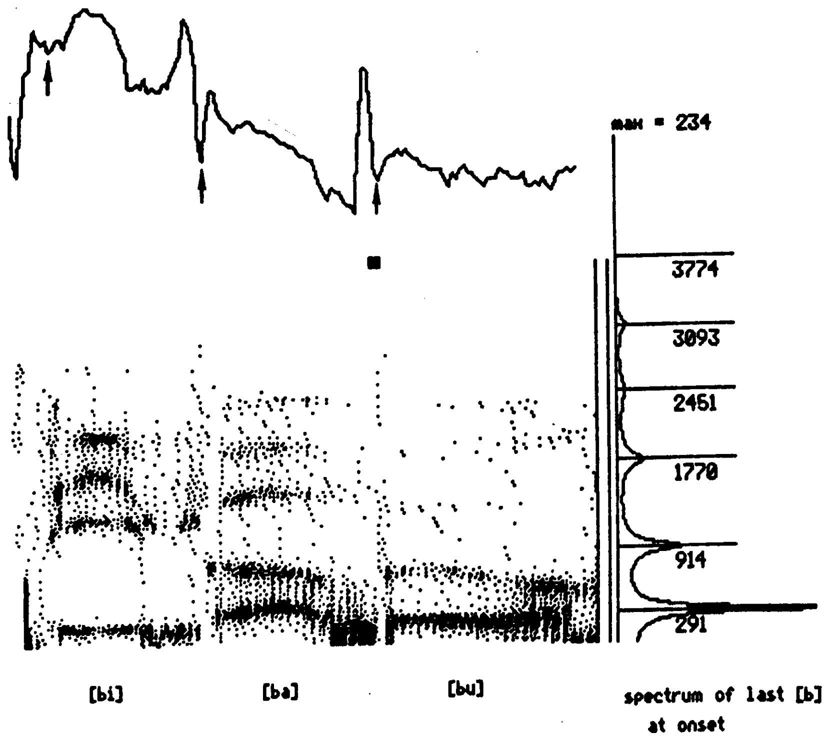

FIG. 17.4. Spectrograms of the syllables [bi], [ba], [bu]. The trace at the top indicates the time-changing values for ACUTEness of the signal. Arrows indicate the moment of release for the three bilabial stops.

previous work, we had initially hoped that these continua would be less context sensitive than the others. However, even the values of the LOCUS parameter as extracted by our preprocessor vary lawfully from context to context.

This lawful variability is dealt with in TRACE in a relatively direct manner. Quite simply, phoneme nodes have connections which modulate the strengths of the interconnections between feature nodes and phoneme nodes in adjacent time slices. An illustration of this is shown in Figure 17.5. Here we see connections between ACUTE feature detectors and one of three stops. (Again, for simplicity we show only three of the eight feature detectors, and we omit the connections between each of these detectors and the other two stops.) The actual values for ACUTE which are observed for each stop are heavily influenced by the vowels which occur later in time. So, phoneme nodes for the vowels [a], [i], and [u] in slice eight are allowed to modulate the weights of the feature-to-stop connections. In essence, the set of weights is the sum of the sets of weights appropriate for each vowel. The magnitude of the contribution made by each vowel-appropriate set is determined by the activation level of the detector for that vowel in surrounding slices. Thus, if only one vowel is active in the context, the feature to phoneme weights will be dominated by the weights appropriate for the consonant in question when it occurs with that vowel. If several vowels are partially active, the feature to phoneme weights will reflect a composite of the weights appropriate for each vowel. When no vowels are active, the weights are a composite of the sets appropriate for each vowel.

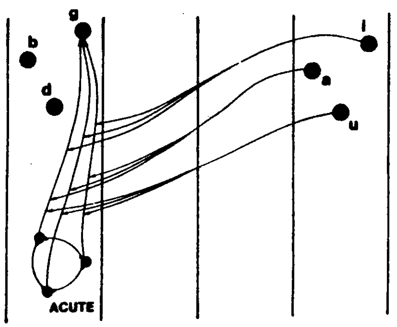

FIG. 17.5. An example of how contextual effects are dealt with in the model. The phoneme nodes for the vowels [a], [i], and [u] are able to modulate the strength of the connections between ACUTE feature nodes and the nodes for the stops [b], [d], and [g]. (This is shown here only for the connections to the velar stop.) This allows the model to compensate for the effect that the vowel has on the spectrum of the stop.

The weight-modulation scheme represents an important increment in the computational capabilities of the nodes, relative to earlier interactive activation models. Not only can they excite and inhibit other nodes, they can also determine how strongly one node can excite or inhibit another. In general, this is an extremely powerful addition to interactive activation models because it allows them to behave in a more sharply context-sensitive fashion than might otherwise be possible. The idea of modulating the weights was introduced by Hinton (1981), and is now being applied to a number of different problems in which the mapping from one level of description onto another is highly context dependent.

SIMULATIONS AND EXPERIMENTS

The TRACE model has been implemented as a program written in the C programming language on a VAX 11/750 digital computer. There are acually two versions of this program; one accepts real speech input and the ther processes simulated (speech-like) input. The first version is used to test he operation of the feature and phoneme interactions, and the second, not onsidered here, is used to explore the properties of the phoneme and wordevel interactions. Here we present only the results of simulations using the eal-speech model, focusing on the way in which the TRACE model exploits awful variability by using variable weights to improve its identification of honemes in context.

ontext Effects on Stop Identification

Ve have tested the model by presenting it with 25 tokens of each of the yllables [ba], [bi], [bu], [da], [di], [du], [ga], [gi], [gu]. These tokens were ttered by the same speaker whose speech was used to tune the connections etween the feature detectors and phoneme nodes, but the test tokens were ot part of the training set.

We have tested the model under two different conditions. In one, called he "fixed weight condition," a fixed set of weights were used linking the ature and phoneme levels. In the other, called the "variable weight conition," the weight modulation scheme described above was used. In both ses, the weights were calculated using the perceptron convergence proceure (Rosenblatt, 1962). For the fixed-weight condition, the training data r each target phoneme included instances of all three vowel contexts. For e variable-weight condition, the training was carried out separately for ch of the three contexts. Thus, the best compromise set of weights that uld be found using the perceptron convergence procedure was used in the xed-weight condition, while for the variable-weight condition, the best set weights was found for each of the three vowel contexts.

In Figure 17.6 we see the results of presentation of one of the [ba] tokens the model. The columns represent every phoneme slice, indexed by the ature slice it is centered over. Time begins at feature-slice 0 and continues feature-slice 57. Since each feature slice lasts 5 msec, the columns display tivity occurring at 15 msec intervals.

Of interest is the changing pattern of activation of the phoneme nodes. e show this by using letters to designate phonemes whose activations have ssed threshold; the height of the letter in the column indicates the relative tivation of that phoneme node. Because activity within a time slice contins even after the input is directed to later slices, we can look at the state of e trace at several different points in time. Figure 17.6 only shows activity the points in time when the input is directed at slice 6 (30 msec) and slice (105 msec). As mentioned previously, the TRACE is both perceptual presentation and working memory, in that it contains a record of the past ck to the beginning of the trace which can continue to evolve as it processes

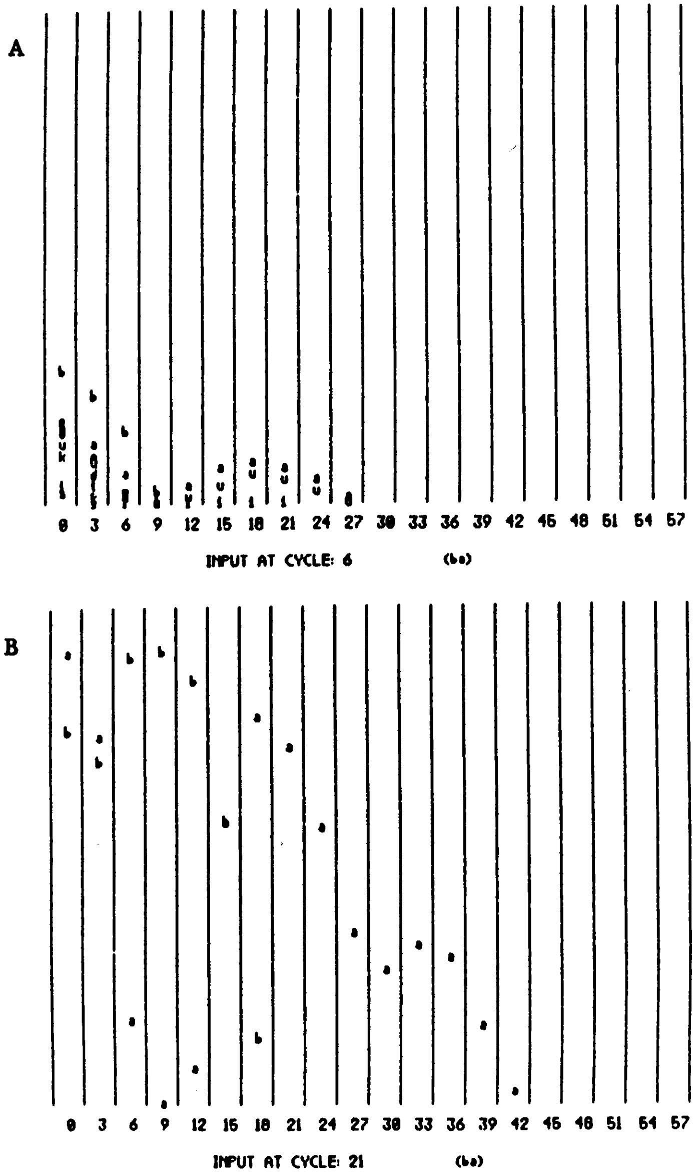

FIG. 17.6. Activation levels of phoneme nodes at 30 msec (a) and 105 msec (b) with presentation of [ba]. Vertical columns correspond to 5 msec time slices; every third slice is shown. The activation level of a phoneme is shown by its height within a column.

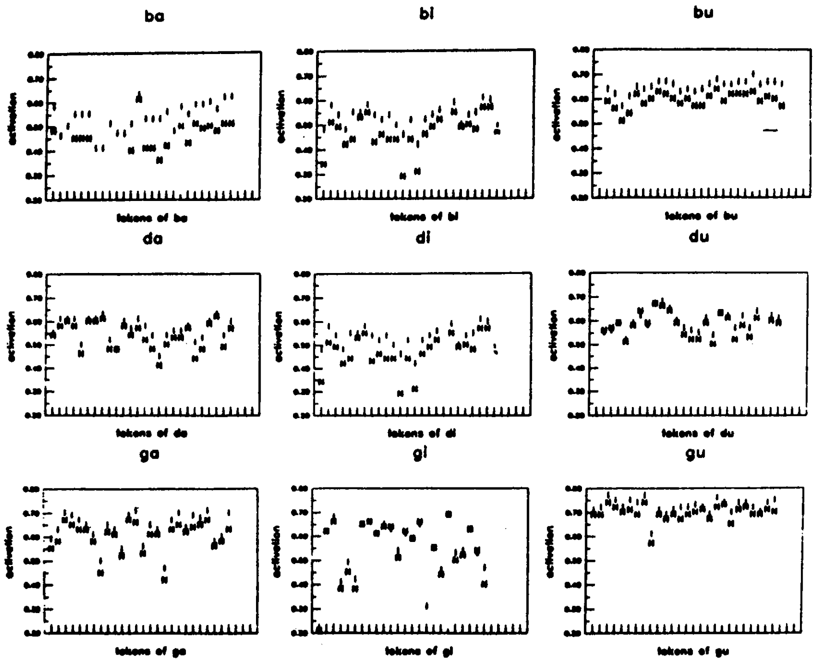

FIG. 17.7. Results of presenting 25 tokens of each of the syllables [ba], [bi], [bu], [da], [di], [du], [ga], [gi], [gu] to the model. Results for each token are shown along the horizontal axis; activation level is given along the vertical axis. The letter *I* marks activation levels when context effects are handled by node interactions.The letter *N* indicates results when interactions are turned off.

information arriving further down the trace. There is no sharp distinction between the structures which are responsible for on-going perception, and those traditionally called short-term perceptual "memory."

All phonemes are below threshold before the input is presented to TRACE. In Figure 17.6a we see that by approximately 30 msec after release of the beginning of the sound, the bilabial stop has become quite active. We also see that there are vowels which have become active in later traces—traces which in fact have not yet received speech input. The system is able to anticipate the presence of the vowels because in fact the speech that is "heard" in Trace 6 contains strong cues as to vowel identity. By 105 msec (Figure 17.6b) the [b] has become clearly established, and the identity of the vowel is clear. Note that although there are clear "winners," at no point is TRACE forced to make a definitive decision about the nature of the perceived

sounds. Perception is treated as a graded phenomenon, and it is possible for the perceptual system to entertain several candidate interpretations for the same stretch of input simultaneously.

The results shown in Figure 17.6 describe the time-course of activation of phoneme nodes for a single token of the syllable [ba] in the variable weight condition. Summary results of the simulation runs from both conditions are shown in Figure 17.7. The nine panels show, for each of the syllable types, the level of activation attained at timeslice 35 (125 msec) by the target stop node in the variable mapping condition (labeled "I") and in the fixed mapping condition ("N"), for each of the 25 different speech tokens used in each condition. In nearly every case these interactions were of at least some benefit, increasing the activation level achieved by the correct target phoneme. There are some syllables where the difference is not very great, ([du], [gi]) but even here there is an advantage for the variable weight condition, particularly for those tokens which do the most poorly in the fixed-weight condition. (In cases where a column contains no N, the activation of the correct phoneme was below .2, the minimum value shown in the figure).

Table 17.2 shows the percentage of different tokens of the stop-vowel pairs in which the correct consonant was more strongly activated than either of the other two, for the fixed and variable mapping conditions. In all cases, TRACE's performance was better with the variable mapping in place. In nearly every case, performance was near perfect with variable mappings. When the context was allowed to modulate the weights, performance was 90% correct; when the context was not allowed to modulate the weights, and the fixed set of weights was used, performance was only 79% correct. The only syllable which presented serious problems was [bi].

It should be noted that the fixed set of weights used in the fixed weight condition was not the average of the weights used for the three different vowel contexts. It was the best set we could find to discriminate between the classes of sounds when the model was forced to find one set of weights to use in all contexts, given the feature set we were using, and the scheme of weight determination described above.

Of course, it might be possible to do better with a different feature set and/or a different procedure for choosing the weights. Whenever there are invariant features, it would seem likely that the perceptual system would exploit them, and if researchers find such features we would certainly want to incorporate them into our model. Our point is simply that the weight modulation scheme employed here provides a way of exploiting lawful variability which cannot be captured in invariant features. Our simulation simply shows that, given a particular set of features, the addition of weight modulation can result in improvements in performance.

TABLE 17.2

Number of Correct Identifications in 25 Trials
Variable Weight and Fixed Weight Results
(Percentage Correct)

[ba]		[bi]		[bu]	
variable	fixed	variable	fixed	variable	fixed
25 (100)	17 (68)	14 (56)	13 (52)	25 (100)	20 (80)
[da]		**[di]**		**[du]**	
variable	fixed	variable	fixed	variable	fixed
24 (96)	21 (84)	21 (84)	20 (80)	23 (92)	21 (84)
[ga]		**[gi]**		**[gu]**	
variable	fixed	variable	fixed	variable	fixed
22 (88)	21 (84)	24 (96)	22 (88)	25 (100)	24 (96)

Overall percentage correct for variable weight condition: 90
Overall percentage correct for fixed weight condition: 79

Context Effects on Vowel Identification

Much has been made of the lack of invariance in acoustic specification of stops, but it is also true that vowels sometimes exhibit great variability as a function of their context. During the course of building the feature detectors, we became aware of environments where this was particularly true. In Figure 17.8, we see running spectra of three tokens of the vowel [u]. These tokens were excised from the syllables [bu], [du], and [gu]. In each case they begin 150 msec after the release of the stop, which is beyond the point where the obvious consonantal transitions have occurred. The clear differences in these spectra impelled us to modify TRACE in a way for vowels which resembled what we did for the stops. We permitted stop nodes in early traces to interact with certain connections between nodes for features and vowels in later traces (this was allowed to happen only for those feature-vowel combinations where we observed there to be significant context effects). As was the case with the vowel node effects on feature-stop interactions, the modulations were adjusted to compensate for the observed contextual effects.

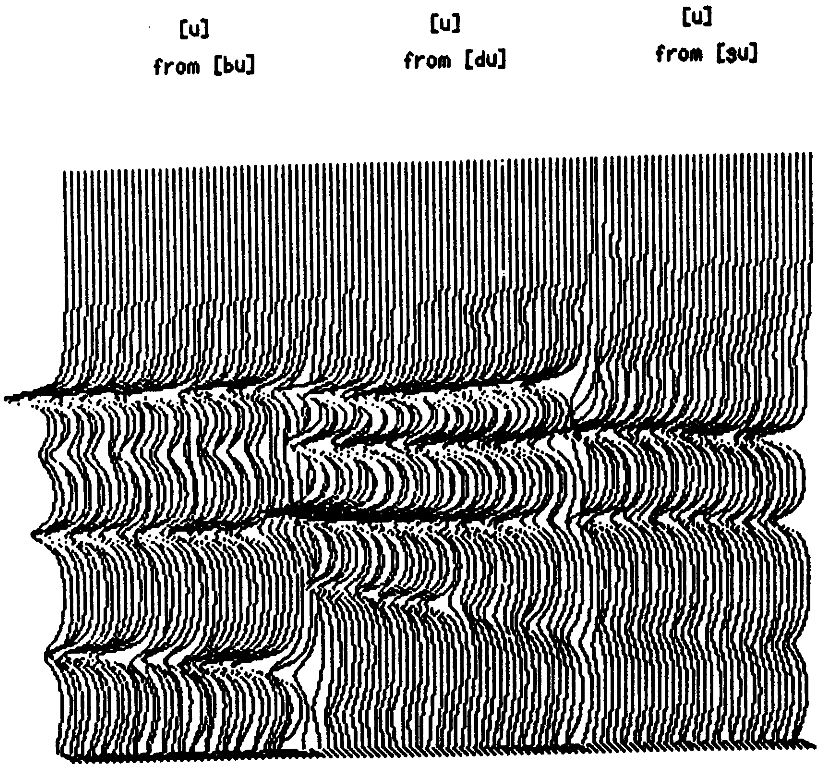

FIG. 17.8. Running spectra (at 5 msec intervals) of the vowel [u], excised from [bu], [du], and [gu]. The initial 200 msec of each syllable was removed, deleting all of the consonantal formant transitions. Spectra are noticeably different.

These interactions provide important information to the model, and illustrate how contextual effects may be exploited in the recognition process. Figure 17.9 shows how the model responds to input consisting of the final 200 msec of a [gu] token; the initial 175 msec were deleted. Although perceives a vowel at all points in time, during the first few traces the stop node [g] has become somewhat active.

Presumably, human listeners would be able to take advantage of this information as well. To test this, we presented subjects with a series of vowels which had been extracted from the final 200 msec of CV syllables which

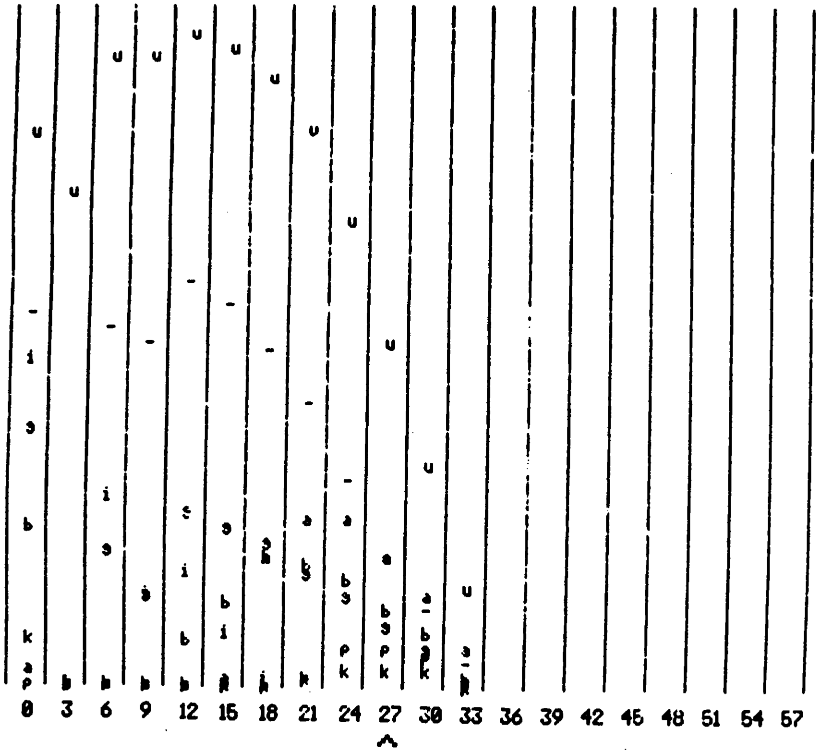

FIG. 17.9. Activation levels of phoneme nodes at 135 msec with presentation of [u] excised from [gu] context. Although the [u] node dominates, evidence for the context is present and the [g] node is highly activated.

combined the stops [b], [d], and [g] with the vowels [a], [i], and [u] (the initial 175 msec were deleted). Subjects were told that they would hear a series of vowels which had been recorded in a CV context, with the C digitally removed. They were asked to guess what the missing C was. The confusion matrix of subjects' responses for the vowels extracted from syllables [bu], [gu], [du] is presented in Table 17.3. We see that subjects often able to recover the identity of the missing consonant. These preliminary data provide nice support for the notion that TRACE's attempt to take advantage of variability—rather than trying to ignore it—corresponds to the approach human perceivers take.

TABLE 17.3

Response Probabilities for Identifying Preceeding Stop from Steady-state Portion of the Vowel

	Original syllable		
Response	*bu*	*du*	*gu*
b	.61	.01	.33
d	.10	.97	.24
g	.28	.01	.43

Note: Response probabilities are based on 70 observations for each vowel.

WHEN EXPLOITATION OF LAWFUL VARIABILITY IS NOT ENOUGH

The speech stream contains more information than most approaches successfully exploit, but sometimes it does not contain enough to specify the correct sequence of phonemes uniquely. In these cases, the model's ability to exploit top-down input from the word level to the phoneme level causes the model to prefer lexically acceptable alternatives consistent with the input to lexically unacceptable alternatives. In a forthcoming paper (McClelland and Elman, in preparation), we show how the simulated-speech version of the TRACE model can identify words and use word-level activations to bias phoneme-level activations. Strikingly, the model does not require any word-boundary information. It can pick out the sequence of words from the pattern of phoneme-level activations (where this is not ambiguous), just as the phoneme level picks out the sequence of phonemes from the pattern of feature level activations.

CONCLUSION

Our view on invariance in speech differs from that of some of the other contributors to this volume. We see the lack of acoustic invariance in the speech signal as a positive thing. This variability provides listeners with a rich source of information. We do not deny that it may be possible to transform the acoustic data so as to recover some properties which remain relatively invariant across contexts. But we doubt that this will be true of all aspects of the signal, or that many transformations can be carried out without regard for the context. What we hope we have demonstrated here is that the interactive activation framework embodied in the TRACE model provides a powerful way to process complex contextual interactions.

Hopefully, it also provides greater insight into the ways in which human listeners perceive speech. Despite the simplicity of our distinctive features, TRACE's overall level of performance on the perceptual identification tasks was 90%.

An important reason for the success of the TRACE model is the basic architecture of the TRACE itself. The TRACE allows activations at different points in time to exert mutual influences. Both context which comes before a section of the input, and context which comes after it, can influence the final configuration of activations over the section of the input and therefore the interpretations which that input receives. Local coarticulatory influences of one phoneme on another can be handled by weight modulation, while constraints arising from the lexical and, of course, higher levels not yet implemented can be handled via top-down excitation. Thus, the model is capable of filling in a missing phoneme at the beginning of a word based on the rest or the word, and with higher levels would be able to exploit information from subsequent words. Models which work in a more strictly left-to-right mode cannot cope as easily with backward as well as forward contextual influences, without invoking reprocessing. Yet is is clear that backward contextual influences are crucial if we are to account either for the effects of local phonetic context or lexical and semantic context on phoneme identification. Experiments by Isenberg, Walker, Ryder, & Schweickert (1980) have shown that such backward effects can extend to the preceding word.

We also note that the model provides a solution to a problem which has been troublesome to many recognition schemes. Typically, one of the first tasks undertaken by models of speech recognition is the segmentation of the input. This is made difficult by the frequent lack of obvious segment boundaries. The process is usually fraught with errors which are then compounded at later stages of processing. In contrast, the TRACE model does not attempt an explicit parsing of the input into segments. It recognizes that segmental information is blended across broad stretches of input. Nonetheless, while recognizing the essentially continuous nature of the acoustic input, the model does maintain a level of representation which consists of discrete units (phonemes); it is simply the case that the detectors for these inputs do not require a segmented input, but simply respond to the appropriate pattern or something close to it wherever it occurs in the speech stream.

ACKNOWLEDGEMENTS

This work was supported by contract N00014-82-C-0374 from the Office of Naval Research. We wish to thank Thomas H. Ward for help in development of the signal processing software used by the model, and Paul S. Smith for assistance in carrying out experiments.

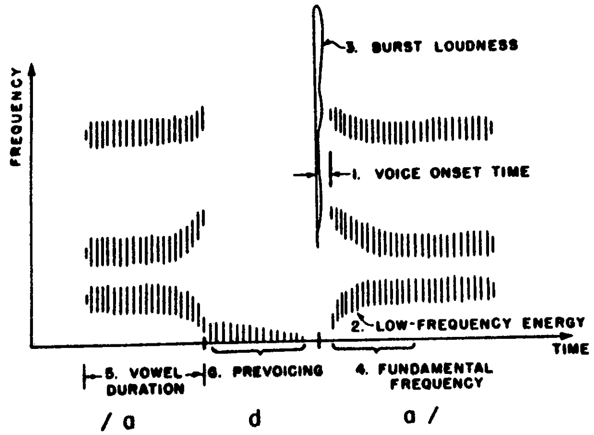

FIG. 17a.1. Analysis of the [d] phoneme in [ada].

Dennis H. Klatt: Comment

A spreading-activation model of speech perception has a great deal of attraction to me because of my interest in artificial intelligence models that began in graduate school many years ago. This is the most ambitious, most specific, and most powerful of the attempts to use neural analogies to formulate a realistic model of auditory perception. Yet there are interesting ways in which the model appears to me to be totally wrong. Consider for example the problem of recognizing the [d] phoneme in the utterance [ada] shown in Figure 17a.1.

There are many cues that contribute to the decision, and they are distributed over a considerable interval of time. For example, to decide whether the consonant is a [t] or a [d], one has to consider at least six more-or-less distinct cues, ranging from the duration of the preceding vowel and the extent of the VC formant transitions to the presence of voicing energy during closure, burst intensity, presence of aspiration, amount of low frequency energy present at voicing onset, and CV formant transition extent.

In the model of Elman and McClelland, there are a set of acoustic feature detectors that, by and large, examine what is going on locally—which appears consistent with the view implied in Figure 17a.1 that these six separate cues must be detected and then integrated in the decision process. However, there are problems in dealing with the natural variability in the durations of acoustic events such as those shown in the figure.

<u>Problem 1</u> is to compute the value of a cue over the proper interval of time. Take for example the detector whose job it is to detect the presence or absence of a voicebar (cue 6 in Fig. 17a.1) over some interval of time. If the closure interval is short, as it may sometimes be, and the detector is simply looking for low-frequency energy over a fixed predetermined time interval, energy from the vowels may falsely indicate a voicebar in what is really a voiceless plosive. Elman and McClelland propose

to solve this problem by postulating multiple feature detectors, each examining a time interval of different length or different location. If the closure interval is long, then one instantiation of the detector, the one attending to the time interval just after closure, has the right answer at its output, but how does the next level of feature integrating phoneme detectors know which of the multiple instances of each detector to believe?

Problem 2 is to put together the information distributed over time when the temporal location of each cue can vary by as much as it does in normal connected speech. It is well known that a vowel such as [a] can vary in duration over a range from 50 to 300 ms or more depending on the postvocalic consonant, syllable stress, and position of the syllable in the word and in the phrase (Klatt, 1976), and that the duration of a consonant such as [d] (excluding flaps) can vary from 30 to 150 ms or more. How does one then integrate the VC transition cue with the CV transition cue without requiring each detector to say, in effect, "I am giving the best answer of all the VC detector instantiations scattered along the time axis"?

Expressed in another way, problem 2 boils down to an inability in this type of model to treat time as an independent dimension[1] in order to note the *sequence* of acoustic events and then ask if the event sequence is one expected for [d]—only secondarily asking if the events are separated in time by complex relationships unique to [d]. It is exactly this idea of gremlins monitoring the input for a specified sequence of acoustic events required to see [d] and other phonetic segments that led me to formulate Scriber (and then LAFS) as plausible functional models of speech perception (Klatt, 1979). Until a spreading activation model can perform computations similar in kind to those of Scriber and LAFS, I do not believe that it can be taken seriously as a realistic model of perceptual processing.

Problem 3 is that many kinds of context sensitivities and cue trading relationships known to affect perception are not possible to realize in this type of model. One can approximate virtually any tendency in individual cases, which makes it hard to document this as a serious deficiency. However, I believe that the indirect way in which knowledge of phonetic context must be incorporated by a weighting scheme is a hint that some perceptual peculiarities ought to exist, i.e. some relationships would be unlearnable in the sense the perceptrons are known to be limited in the kinds of generalizations that they can make—whereas listeners seem to know everything that is regular in speech acoustic phonetics.

Jeffrey L. Elman, and James L. McClelland: Response

Klatt is correct to point out that the way in which the speech signal varies with changes in speech rate poses important challenges to our approach. However, we believe that the basic interactive activation framework, and in particular the mechanisms we have described for exploiting lawful variability, can meet the challenge. The fundamental point is that the variation in the speech wave as a function of rate and stress is lawful. Just as we were able to exploit this lawful variability by modulating the pattern of connections between feature and phoneme levels as a function of local phonetic context, so we anticipate we will be able to modulate these same connections as a function of on-going speech rate, and even momentary changes in rate. In principle such a mechanism is more powerful than the time warping presently preferred, because it treats the temporal structure of speech as more than just a succession of events at unspecified intervals. There is regularity in the exact details of the way in which speech compresses (and distorts) as it is produced with greater and greater haste, and we believe our approach will be able to exploit this regularity quite effectively. We hope to report progress in demonstrating these points in the near future.

[1] One can perhaps devise counters to measure the length of vowels or other events.

Leigh Lisker: Comment

We seem generally to be in thorough agreement as to what the identity of a word is, while we worry about what the phoneme is; at the same time, we usually think that we need the phonemes in order to decide what the word is. In the oral presentation of their paper, McClelland mentioned a hypothetical case of "barricade" and "parakeet" which is of interest when we consider the question of phoneme perception. I am struck by the expression he used: "along comes the *icade*, so you know that it was a [b], and if it was *akeet* you know it was a [p]." This situation suggests that we decide what the phoneme was on the basis of having identified the word. We tend to think of the phonemes as known, on the basis of which we want to explain how we manage to get the words of a speech message; McClelland's way of looking at the matter of speech perception is in some sense precisely backwards.

Jeffrey L. Elman, and James L. McClelland: Response

Lisker's comment reflects a vision of the perceptual mechanism which predates the kind of model we are proposing. Models conceived within the context of this older vision viewed phoneme-level processing as a process which assigns identities to phonetic segments in an all-or-nothing fashion, in a sequence of discrete events. In such models, it is of course impossible to see how phoneme identification can be influenced by word recognition if the word is to be recognized on the basis of the output of this very phoneme-identification process. However, our model does not share this implicit assumption that phoneme processing is a series of discrete events. Instead, we see the process of forming representations at the phoneme level as one that takes place slowly and continuously in time, with partial information being passed on to other levels as it is available, and with information being assimilated from other levels as they make information available to the phoneme level. Thus, it is perfectly plausible and reasonable to assume that phoneme level processes can make available information such as "the first phoneme is [p] or [b]; the second [u] and the third [l]" to the word level; and for the word level to use this information to decide that the string must begin with a [p] (since no words in English begin with [bul]). More generally, Lisker appears to be confusing two issues: Whether a process is influenced by information received from another (contingency), and whether a process can only start once another has finished (seriality). Over the years, this assumption has often been held in a number of fields, but it is far from necessarily true. Indeed it is possible for processing to be parallel and at two different levels, and for each level to be contingent on the other, as long as it is possible for each to share partial tentative results with the other. Mechanisms which use this type of parallel-contingent processing are capable of exploiting mutual constraint between different processing levels quite effectively.

O. Fujimura: Comment

The use of phonemes as the basic segmental units for this scheme is not the best choice. Basically, one would like to have a concatenative string of units as the abstract representation of the signal in such a way that the inventory of units (nodes) to be compared at each time sample represents a truly competitive paradigm. Often, however, a time slice represents a syntagmatic combination of phonemic segments, for example, a consonant plus vowel. The consonant and the vowel in such cases (not only in transition, but also in quasi-stationary parts of the signal) do not compete with each other, as the candidate for that slice. Moreover, a noncontiguous interaction (between syllable nuclei ignoring intervening consonants, for example) must be handled. While the network manipulation *can* handle such intersegmental interactions and constraints, it tends to make the framework excessively powerful and computationally intractable. If the phonological units themselves are chosen more naturally for the paradigmatic phonetic structures of speech, the scheme should work more efficiently and algorithm should be substantially simpler. I recommend the use of demisyllables (Fujimura, 1979; Fujimura and Lovins, 1978) as better units for this purpose (of course, at the cost of more units in storage, which I think is justifiable). Then, most of strong and often *ad hoc* intersegmental interactions are already built in in the stored patterns, and noncontiguous interactions can be handled by general interactions between values of specific features (such as vocalic features of contiguous, perhaps only stressed, syllables).

Michael Studdert-Kennedy: Comment

Assessing the perceptual value of coarticulatory effects may be a delicate matter, as the model of Elman & McClelland suggests. For example, Lehiste and Shockey (1972) found that listeners to a VC fragment, cut at the closure of a V_1–stop–V_2 disyllable, could not identify the vowel that had followed in the original utterance. However, if we want to characterize subtle properties of the information in the signal as it accumulates to determine a response, it may not be enough simply to see whether the response occurs or not. Several experiments by Martin and Bunnell (1981, 1982) using [(C)VCV] spondee frames, have shown that reaction time increases and accuracy of indentification decreases for the second vowel, if it has been cross-spliced to follow a context other than that in which it was spoken. Alfonso and Baer (1982) found that listeners can often identify the final stresses vowel in disyllables of the form [əp V p], from information in the initial unstressed vowel alone. Both these studies also showed systematic and predictable acoustic differences in preceding vowel as a function of following context. Finally, Fowler (1981) showed that listeners can discrininate among medial stressed vowels, gated from trisyllables of the form [V b ʌ b V], as a function of the context in which they were originally spoken. (See also, Chapter 6.) None of these results could have occurred if listeners were not sensitive to acoustic changes reflecting articulatory adjustments within one syllable preparatory to execution of the next.