Psychological Bulletin
1999, Vol. 125, No. 4, 410–436

0033-2909/99/$3.00

Connectionist Modeling of Speech Perception

Athanassios Protopapas
Brown University

Connectionist models of perception and cognition, including the process of deducing meaningful messages from patterns of acoustic waves emitted by vocal tracts, are developed and refined as human understanding of brain function, psychological processes, and the properties of massively parallel architectures advances. The present article presents several important contributions from diverse points of view in the area of connectionist modeling of speech perception and discusses their relative merits with respect to specific theoretical issues and empirical findings. TRACE, the Elman/Norris net, and Adaptive Resonance Theory constitute pivotal points exemplifying overall modeling success, progress in temporal representation, and plausible modeling of learning, respectively. Other modeling efforts are presented for the specific insights they offer, and the article concludes with a discussion of computational versus dynamic modeling of phonological processes.

Connectionist modeling, a term often used synonymously with neural network modeling, refers to a class of models with a special, intrinsically parallel, architecture. These models consist of a number of interconnected units, or nodes, with modifiable connection weights, which determine the strength of influence one node can exert on another. Such models have been used in many areas of psychological modeling with varying degrees of success. In this article, the psychological process of interest is speech perception, that is, the transformation of the acoustical speech signal to meaningful lexical items (words or morphemes). A brief introduction to the basic issues in speech perception and word recognition is presented first. Some key concepts and terms of connectionist modeling are also explained in the introduction. In the following sections, particular approaches to speech perception modeling are reviewed, conceptually organized around the major issues that remain to be solved. Modeling efforts are evaluated on grounds of psychological validity and biological plausibility and not by their performance in task-limited situations. Empirical findings from experimental psychology are brought to bear on the discussion whenever appropriate.

Basic Issues in Human Speech Perception

In the present article, the term *speech perception* is meant to encompass the processing performed by the human brain that begins with the auditory registration of a speech waveform and concludes with the identification of the spoken words in it. The end goal in speech communication is the registration of the meaning of entire utterances and not merely the identification of their individual meaning-carrying constituents. Nevertheless, ignoring the complexities arising from syntactic and semantic processing is a common initial approach, warranted to the extent that a good deal of progress can be achieved by breaking down the complete problem into more manageable pieces. By this working definition, it becomes clear that researchers need to understand the way information is represented at the various stages of processing from acoustic to lexical, as well as the processes that perform the mappings. In the following paragraphs, key concepts and problems in speech perception modeling are introduced, necessarily greatly simplified because of space considerations. Presentation of each theme is made with the goal of orienting the reader to be better able to evaluate the models presented later, and this is obvious in that only those points of view that most easily lend themselves to connectionist conceptualization are developed. For more balanced presentations and thorough discussions (and references) of the basic issues in speech perception, one is advised to consult, for example, Altmann (1990), Altmann and Shillcock (1993), Marslen-Wilson (1989), and Miller and Eimas (1995a, 1995b). In the discussion of specific models, selected empirical findings are introduced when they are necessary for understanding and evaluating the claims of and arguments for and against each model.

Conventional wisdom holds that some form of power spectrum (i.e., the distribution of acoustic energy across frequency bands over time) constitutes the representation of sound that enters the auditory cortex (see, e.g., Blomberg, Carlson, Elenius, & Granström, 1986, and following commentaries). Most modeling efforts to date, as well as engineering attempts at speech recognition machines, ignore the complexity arising from the existence of sounds other than clear speech from a single speaker in this power spectrum and concentrate on the processing of a single speech stream. This is the case in all the models reviewed below, and, though a special and (probably) unusual case, it is implicit in all following discussion.

Assuming an initial auditory representation of a single speech stream, how are words identified in it? The auditory representation

Athanassios Protopapas, Department of Cognitive and Linguistic Sciences, Brown University.

Preparation of this article was supported in part by Grant HD 05331 from the National Institute of Health and Human Development to Peter D. Eimas. Many thanks are due to Peter Eimas, Paul Allopenna, Philip Lieberman, James Morgan, James Anderson, David Ascher, Gary Strangman, and David Landan for helpful and stimulating discussions and comments on earlier drafts of this article.

Correspondence concerning this article should be addressed to Athanassios Protopapas, who is now at the Department of Psychology, Panteion University of Athens, Leoforos Syngrou 136, GR-176 71 Athens, Greece. Electronic mail may be sent to protopap@panteion.gr.

must be transformed into the form in which the lexicon is encoded before a matching process can take place. Because of the great variability in the acoustic realization of lexical items, discussed below, directly mapping from acoustic to lexical representations is considered at best impractical. That is, the lexicon cannot be specified in terms of raw power spectra, and thus prelexical representations are commonly hypothesized, that is, transformations of the sound properties into representations of a linguistic nature, before contact is made with the mental lexicon. The prelexical representations most frequently considered are features, phonemes (often called plainly *segments*), and syllables.

The term *feature,* often preceded by the qualifier *acoustic/ phonetic,* is used here to refer to acoustic characteristics (and constellations thereof) that are useful (though not necessarily distinctive) in a language because they correlate with distinctions between classes of speech sounds and, therefore, with differences in meaning. Sometimes a notion more akin to phonological features is used in modeling, referring to abstract properties of an idealized speech representation that need not have any acoustic or articulatory physical correlate and only serve to distinguish between abstract phonological categories that are used to describe phonological phenomena. The particular meaning of the term *feature* in a model of speech perception tends to depend mostly on the distance between a model's input representation and real speech (i.e., an acoustic signal).

It seems to be an implicit (if unwarranted) assumption in much of the speech literature that the speech signal is transformed to a phonemic representation prior to making contact with the mental lexicon. The term *phoneme* is used in phonology to refer, at an abstract level, to constituent parts of words. Phonemes are often represented by letters in alphabetic scripts; for example, the word *bat* is made up by the phonemes /b/, /æ/, and /t/, spelled with the letters B, A, and T, respectively. Elegant and efficient descriptions of the phonological organization of language are made with reference to phonemes. In acoustic phonetics, however, phonemes remain elusive; that is, it has not been possible to identify invariant acoustic correlates of phonemes in the speech stream. If it were possible to decode the speech waveform into a series of phonemes, a large part of the problem of speech perception would be solved. In reality, the variability in acoustic realization of phonemes (assuming, for a moment, that phonemes are the fundamental blocks of speech communication) restricts our options to classifying classes of sounds, often termed *phones,* into phonemic categories; this is by no means a trivial task. With respect to the treatment of phonemes, a sharp distinction between models is evident: On the one hand, there are models that attempt to explicitly identify phonemes from the speech stream on the basis of acoustic properties. The scope of these models typically excludes word recognition. On the other hand, there are models that take phonemic representation for granted in their input and are concerned with presumably subsequent processing stages. Unless great advances are made with respect to the role of phonemes in speech communication, neither approach is likely to prove particularly fruitful when it comes to integrating acoustic processing with lexical access.

The third candidate unit for prelexical representation is the syllable, which, in phonology, is a unit of organization of ordered series of phonemes into language-specific structures. It has been argued that syllables also constitute the unit of motor planning in speech production, and recent empirical work has examined the role of syllables in perception. So far, syllabic conceptualization has depended heavily on phonemic representations, that is, the syllable is viewed as a structured set of phonemes rather than as a unit, or as a vehicle for encoding and coproducing phonetic features that may perhaps later be perceptually assembled into phonemes or directly into words. There is little mention of syllables in the models discussed below; however, implicit reliance on syllable-sized units is sometimes evident in the form of phonotactics and context-dependent phonemes. Specifically, phonotactics are language-specific constraints on the possible phoneme sequences that are usually defined with respect to a syllabic frame. Context-dependent phonemes are a practical way of retaining the phoneme as the basic unit while acknowledging its contextual variability, but it may be hard to distinguish this from a syllabic representation to the extent that syllabic position constitutes a primary correlate of acoustic variation.

Acoustic variability in the realization of lexical items includes variability in the physical properties of the sound sources (e.g., differences in size between the larynges and the vocal tracts of different speakers) as well as within-speaker variability arising from differences in the situation in which an utterance was produced (e.g., ambient noise, emotional state of the speaker, speaking rate). In addition, depending on one's choice of prelexical units, there is variability in the acoustic realization of these units arising from the speech context in which they are produced, generally termed *coarticulation.* For example, the [s] is pronounced differently in *sue* and in *see* because of the influence of the following vowel, which partly determines the vocal tract configuration while the [s] is produced.

An additional source of variability can be found in phonological processes that alter the acoustic realization of lexical items under certain conditions. For example, the word *red* may be pronounced *reg* when the word *car* follows it. More generally, the place of articulation (i.e., the point of maximum constriction in the vocal tract) is, in some cases, assimilated from one phoneme to the one preceding it. In this case (*red car*), the alveolar place of /d/ is changed to velar, to match that of the ensuing /k/. Although listeners have no trouble understanding the intended word (in fact, they may have trouble realizing this phonetic alteration even if instructed to attend to it), it remains a problem for models of speech perception that aim to identify words in the speech stream. Recent modeling attempts have been directed at this problem, as discussed in a later section.

Assuming some form of prelexical representation, it remains to identify the lexical items present in the original speech signal. An obvious first reason this is not a trivial task is that there are no boundary markers in speech, that is, in contrast to the white space that separates words in script there are few (if any) acoustic cues signaling the beginning and end of each word. In fact, coarticulation and phonology apply indiscriminately across words as they do across syllables (recall the *red car* example). Much research and debate has been centered around this issue of segmentation, that is, of defining word boundaries prior to identifying words. An alternative approach has been to sidestep the issue entirely: Just identify all the words possibly present in an utterance and use other processes to sort out which ones make up the best interpretation. For example, given the input /kætəl/, plausibly corresponding to the beginning of *catalog* up to the *l,* it remains possible that the word *cat* has been spoken, followed by *a* and another word that begins with *l*, or that the word *catalyst* is being spoken instead. Or, in a noisy situation, it is possible that the word *cataract* was actually said but the /r/ was heard as /l/. The latter approach would hold that all possibilities are entertained as long as they are not disfavored by subsequent acoustic information or by semantic or syntactic interpretation.

The concepts of activation and competition have been instrumental in shaping our understanding of this sort of problem. Even though it may be possible to define the solution as a serial search through a (possibly structured) set of lexical items until we find the ones consistent with the input, experimental findings indicate that it is fruitful to think of each lexical item as an independent processing unit that gathers evidence for itself and becomes activated as this evidence accummulates (and fading back into its resting state when it mismatches the input). In parallel with this activation process, lexical items compete with each other for stretches of the input: In the above example, *cat* would compete with *catalog* and *catalyst* for claiming the [kæt] portion of the input. Though consensus has not yet been reached, it appears that, at least in the context of connectionist modeling, activation and competition are the most natural ways to think about lexical access. Most current connectionist models of lexical access use some form of activation and competition, either explicit or implicit.

An important controversy in speech perception, as undoubtedly in other mental faculties as well, concerns the directionality of processing. That is, is the flow of information constrained to occur only from sensory registration to subsequent levels of analysis, or does information from later stages of processing flow backwards to affect earlier stages? For example, is phonetic analysis based on properties of the acoustic signal only (and knowledge of phonetics, of course), or does information about known (or possible) words affect how the acoustic signal is interpreted phonetically? The former view, generally termed *autonomous*, is more in keeping with serial processing conceptualizations and sometimes discredited in connectionist modeling. The alternative interactive view holds that there are no in principle limits to the connectivity, hence to the flow of information, within the system. As review of some recent models below indicates, the debate is far from obsolete, as it evolves to challenge our views of what constitutes a processing stage and how much different stages can be reasonably considered to be separate.

One issue often ignored in modeling is the role of learning and development in the formation and continued function of a speech perception system. A great deal of speech perception must necessarily be learned during development, at least to the extent that languages differ in their phonetic properties as well as in their vocabulary. Some processing mechanisms, such as the transformation from acoustic to phonetic and other prelexical representations, can be argued to be innately specified or at least greatly constrained in how they can develop through exposure to a linguistic environment. The particulars of prelexical and lexical representations, however, are necessarily learned, as are all the words in one's language. It is a very important property of the mature speech perception system that new words can be learned with ease without causing old ones to be forgotten. It is also the case that exposure to a word has measurable effects on how subsequent words are processed, that is, the system can be primed as a result of its own processing.

There is now a wealth of evidence that the plasticity of one's speech perception system undergoes substantial changes during one's lifetime. For the purposes of the present review, it suffices to note that infants arrive at their linguistic environment equipped with sophisticated capabilities to discriminate, remember, and group auditory events of linguistic significance. In the first few months of life, the sound system of one's native language is mastered, often at the cost of losing the ability to perceive acoustic differences that are important for making distinctions in other languages. Words and linguistic structures are subsequently mastered with increasing facility as speech production and perception approach their mature, fluent adult targets. Understanding speech perception will depend on discovering the nature of innate biases and the extent to which they shape subsequent development. For this reason, developmental considerations are often invoked below in the presentation of connectionist modeling efforts.

Connectionist Modeling in Psychology

As with my cursory review of key issues in speech perception, space considerations do not permit an adequate presentation of connectionism, the many kinds of models that are available, and the philosophical implications for building and using them. The interested reader is directed to Anderson and Rosenfeld (1988) and Anderson, Pellionisz, and Rosenfeld (1990) for a collection of seminal studies on connectionist modeling, Quinlan (1991) for an introduction to connectionist models in psychological research, and Hertz, Krogh, and Palmer (1991) for a mathematical approach to neural networks. In addition to their contribution to the understanding of innateness in developmental processes, Elman et al. (1996) provide an excellent introduction to modern connectionist thinking and its applications to psychological modeling, with an emphasis on developmental issues.

Inspired by what is known about brain processing structures, connectionist models have been gaining ground in psychology for several decades now. The major strengths of connectionist models, which are often termed *neural networks*, lie in their combination of processing flexibility, massively parallel architecture, and potent statistical generalization. In general, a neural network comprises a number of interconnected units, or nodes, each of which is characterized by an activation value. The connections between the nodes are weighed by the amount in which they allow activation to flow through them from one node to the other. There are no in principle limits to the interconnectivity between nodes; activation can flow simultaneously along any or all connections, and there may be very complex interactions between nodes. The degree of activation of all the nodes and the particular connection weights at a particular instant define the state of the entire model at that instant. Figure 1 (left) shows an example of a generic network of interconnected nodes.

Typically, but not necessarily, the nodes are arranged into layers of particular modeling significance; for example, a subset of nodes may be designated *input nodes*, receiving activation from environmental sources, whereas another subset may be designated *output nodes*, activation of which is taken to constitute the output of the model. Nodes with connections only to other nodes but not to the external world are termed *hidden units* (see Figure 1, middle).

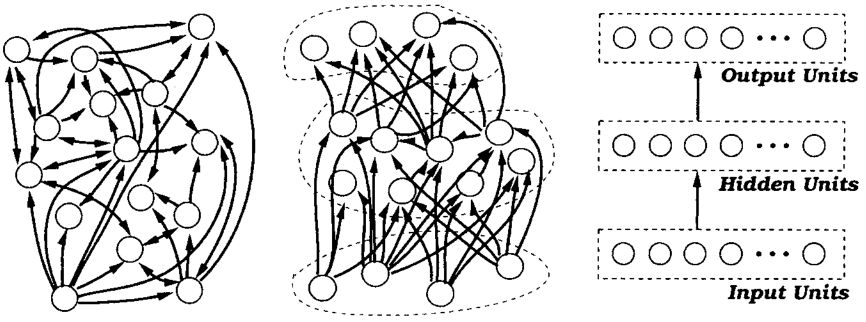

Figure 1. Schematic diagrams of artificial neural networks. Left: General form of an unstructured network of interconnected nodes. Each node, represented by a circle, has an associated activation value at each point in time, and each connection, represented by a line, has an associated weight that controls the amount of activation that may flow between the nodes it connects. The arrows indicate the direction of activation flow between nodes. Middle: Layered network in which groups of nodes, enclosed by dashed lines, are taken to represent levels of information processing of special interest. As indicated by the arrows, the flow of information in this network is restricted to be unidirectional, from the bottom layer to the top layer, and there are no direct connections between nonadjacent layers. Note that connection directionality is independent from node grouping; each reflects an arbitrary decision made on the part of the modeler regarding the type of processing desired in the model and often dictated by practical (e.g., computational) constraints. Right: Shorthand representation of the same three-layer feedforward network, in which arrows between layers signify full connectivity between all node pairs of the two connected layers. Missing connections are equivalent to present connections with zero weight and are used to simplify calculations in matrix form. Dashed boxes enclose an arbitrary number of nodes and are labeled by their functional role in the model as *input, output,* and *hidden* (from the external environment). Training of such networks to map a set of input vectors (i.e., pattern of activation on the input layer of nodes) to output vectors is often done by adjusting the connection weights depending on the difference between actual and desired output for each input pattern. A special algorithm, called back propagation, is used to compute the relative contribution of each connection weight to the final output beginning with the output layer and proceeding backward by taking into account the connectivity and activation levels at each stage.

Such layered construction is very common, and specialized learning algorithms have been developed for training layered models to perform mappings from an input to an output layer. When directionality is thus imposed on a model as a result of specifying an input and an output end, constraints on the flow of information are used to further subdivide classes of networks into *feedforward,* in which activation flows only from input to output without feedback connections (or loops), and *recurrent,* in which there exists at least one closed loop of activation flow.

Connectionist modeling owes much of its appeal to the capacity of the models to learn from a set of exemplars and then successfuly generalize to other similar, but not identical, situations, in a way often strikingly resembling biological behavior. Powerful algorithms have been devised that set the connection weights between nodes to perform a desired mapping between input and output. In some cases a teacher signal is required that modulates the modification of connection weights depending on how closely the actual output of the model resembles a desired output; obviously, the success of such *supervised* algorithms depends critically on defining not only a set of correct input–output relations but also on defining an appropriate metric of what constitutes a good match between an actual and the correct output and a rule for transforming a poor match to a change in connection weights. Gradient descent methods are generally used whereby in each training step the connection weights are updated in the direction in which the output error would be most sharply reduced. Backward error propagation, often abbreviated as *back-prop,* is the most widely used training algorithm of this type for multilayered feedforward networks, updating weights between intermediate layers to a degree proportional to the contribution each weight is computed to have in producing an erroneous output. Alternatively to supervised learning, *unsupervised* learning algorithms operate by exposing the model to a representative variety of inputs and letting the dynamics of the model develop stable representations of the salient features in their input space. Clever arrangement of the architecture of the model and selection of appropriate input stimuli and learning algorithm are critical for the eventual success of a model in either case.

Few issues have attracted as much attention as the notion of *distributed* versus *localist* representation in connectionist modeling. The difference is deceptively simple: In the latter case, each node in a network (or only in part of the network) stands for something of significance to the modeler that can be aptly labeled. For example, there may be a node for the word *boat,* in the sense that activation of that node is taken to directly correspond to the degree of match between the word *boat* and the speech input to the network. This sort of explicit internal representation contrasts sharply with the former, distributed kind, where each node in the network corresponds to nothing in particular but the entire state of the network (or a part of it) is taken to encode what is of interest to the modeler. In general, localist representations are easier to design, understand, and present, and, indeed, they are quite popular in the connectionist literature on psychological modeling. Distributed representations, however, have been often proposed on the basis of more important advantages. Resistance to damage is an oft-cited example: In a localist representation, damaging a node permanently and entirely removes what that node stood for, whereas in a distributed representation damaging any single node may scarcely have any effect. Other important advantages of distributed representations concern the improved signal-to-noise ratio in pattern encoding as well as the more compact representation of patterns and, intrinsically, of the similarities between them.

As mentioned, connectionist modeling is often appealing because of the models' purported resemblance to brain processing structures. In psychology, however, connectionist models are never brain models in any physically meaningful sense (though in neurophysiological modeling things can be quite different). That is, connectionist models are not, in general, brain models; they are at best broad abstractions of brain models. The numerosity of biological networks and the complexity of the chemical interactions that make them functional are well beyond our current modeling reach for any realistic scale. However, the abstraction from a network of real neurons to a connectionist model may be a very reasonable one if it turns out that all the particular, chemical and such, properties of biological neurons are not critical to the function of the network, that is, if the functional properties of large biological networks of real neurons are equivalent to those of much smaller networks of simple abstract interconnected units. It cannot be overemphasized that in psychological models a node is rarely, if ever, meant to correspond to a biological neuron. Rather, by using a parallel network structure that resembles that of biological networks, it is hoped that some fundamental functional characteristics of the entire biological network are retained, even in the absence of close correspondence between the actual processing elements.

Despite this great leap of faith in abstraction, the issue of biological plausibility is not to be brushed aside. Unfortunately, it is often difficult to define specific criteria that can be applied to particular models. In fact, it is easier to identify an implausible aspect of a model than to deem a model biologically plausible. A guiding principle is that connectionist modeling of psychological processes is not about mapping an input to an output with an artificial neural network; it is about understanding how the brain works. Therefore, models may not violate what is known about brain structure; even gross abstractions ought to be justified. A model that incorporates dirty processing tricks to get things done is doomed to irrelevance in the long run. Because the current understanding of brain behavior is rather patchy, concrete criteria are difficult to establish. In the following paragraphs, some examples are given of considerations related to the notion of biological plausibility, to further illuminate its practical aspects (see O'Reilly, 1998, for a more detailed discussion).

There are many ways in which one wants to keep modeling efforts as close to brain function as possible, including constraints on biological learning and processing. For example, if a biological system is known to perform a given mapping within a few processing stages, one would be well advised to construct models of the process that do not exceed the biological process in number of steps (Thorpe & Imbert, 1989). This does not imply that we must always wait for a complete understanding of the biology before we attempt to model the macroscopic scale of events, for not only is such complete understanding never guaranteed, but also modeling work can greatly motivate and inform biological research efforts. Nevertheless, what we know about biology must influence how we build our models; clearly unrealistic architectures and processing schemes are to be avoided.

Moreover, if a biological system is known to constantly adapt to new stimuli in a way that influences subequent processing and adaptation, it is best to build this capacity for on-line learning into the function of the model. Note that a capacity for learning in general is never in itself sufficient for claims of biological plausibility and is only necessary when learning and adaptation are clearly part of the behavior of interest. In this respect, speech perception is not a memoryless process (priming and adaptation effects being cases in point); however, whether adaptive characteristics are critical or not may depend on the particular goals of individual models.

A related issue concerns the type of training that produces a desired behavior in a model and the extent to which an ad hoc learning procedure can be justifiable in a given modeling context. For example, it is one thing to model a more-or-less fixed structure, such as the vestibular–occular reflex, which stabilizes retinal images by causing eye movements that compensate for head movement, or the local bending reflex in the leech (both discussed extensively in Churchland & Sejnowski, 1992, pp. 338–378), and quite another to model lexical activation. In the former cases, evolution has presumably contributed most of the model's structure and function and it is not unreasonable to use any available method to arrive at the connection weights in the model that give rise to the desired behavior, possibly including methods of setting the weights that would be impossible in a biological learning system (e.g., manually setting the weights; cf. the discussion by Churchland & Sejnowski, 1992, pp. 130–139). In the case of lexical activation, however, the function of the model is a result of learning to process the sounds of a language over development. In this case, it is less well justified to count on biologically impossible methods of setting the weights; it would be more informative to explore the kinds of initial biases and learning strategies that result in the mature model than to aim directly for the final product.

Aside from arguments for biological plausibility or psychological parsimony, how connectionist models are evaluated and how they compare with other kinds of models are greatly constrained by the nature of the available evidence and the state of research into the psychological phenomena of interest. Specifically, phonetic perception and word recognition are often investigated in the psychological tradition of setting up experimental conditions, more often than not in a highly unnatural task situation (with respect to normal communication), and establishing the statistical significance of differences between them, with secondary (if any) attention given to quantification (e.g., effect sizes or raw response times). Modeling of such results usually takes the form of a rather qualitative type of data fitting, whereby conditions in the model that correspond to the human experimental conditions are expected to produce the same pattern of differences. In cases where quantitative human data are available (e.g., psychometric functions or time series), a more stringent fit is expected from the model data curves, but the transformations from node activation to the experimental measure can be ad hoc and subject to parameter setting. It is not clear what error measures are appropriate to assess the statistical reliability of a fit and what magnitudes of deviation are acceptable given the greatly simplified nature of models that can be practically implemented. However, all of these criticisms can be levied at least as forcefully against nonconnectionist models, which also generally lack the rich dynamics of connectionist networks that can be examined to provide insights into processes. In addition, connectionist models are by necessity quantitatively implemented. Thus, although implementation constraints may limit their scope, their advantage over qualitative box-and-arrow models is indisputable in that precise and testable predictions can always be generated, sometimes even for experimental manipulations not in the original conception of the model.

It is not an unreasonable expectation that models of speech perception demonstrate, at least in principle, the ability to transform acoustic signals into a representation that is linguistically relevant. This is an area where models are still judged on absolute performance measures, in part because they generally lag so far behind human performance that no rigorous evaluation is yet meaningful, and also because of an interest in automatic speech recognition technology. In the following sections, models of phonetic perception are examined first that learn to extract phonetic properties from the sound signal. The processing of these phonetic properties and their assembly into words is then considered in the next sections, beginning with the TRACE model and continuing with approaches that address its main weaknesses.

Phonetic Categorization

Several researchers have investigated the ability of connectionist models to correctly categorize speech signals into phonetic classes. Most such attempts were driven primarily by performance considerations in the search for an artificial speech-recognizing machine (see Tebelskis, 1995, for a recent review). Some of the findings, however, are relevant in the sense that successful applications provide an existence proof of the power of statistical generalization in the absence of predetermined structures, whereas patterns of failure illustrate the need for specific linguistic constraints in the systems. Demonstrations that phonetic categories may, at least to some extent, be learned solely on the basis of statistical regularities in the acoustic signal, and without reference to articulatory gestures or to innately determined gesture-sound decoding modules, should temper the strongest claims of some nativists who would claim that speech perception, or at least phonetic categorization, is a wholly innate human faculty (e.g., Fodor, 1983; Liberman & Mattingly, 1985).

Training Phonetic Identification

Watrous, Ladendorf, and Kuhn (1990) trained a three-layer network, with recurrent connections from each hidden unit to itself (but otherwise feedforward and without lateral connections), to identify and discriminate between the voiced stops [b], [d], and [g] spoken by a single speaker in the context of a following [i], [a], or [u]. The input to the network consisted of time-aligned spectrogram representations of the syllables. With a different network using delayed connections with simultaneous input from adjacent time frames and a more hierarchically organized structure, Watrous (1990) achieved remarkable performance on the three-consonant forced-choice task in the context of the three vowels encountered during training and perfect performance in consonant–vowel–consonant (CVC) syllables containing the vowels [e] and [æ]. Analysis of the latter network showed that weights were tuned to activate the appropriate units on the basis of spectral characteristics of the consonant release and that the context-related variance associated with each consonant was handled optimally by the system. In the same study, other networks were trained to discriminate consonantal manner of articulation (for a single place of articulation), medial bilabial stop voicing (*rapid* vs. *rabid*), vowels, formant trajectories, and nasal duration, with overall performance always exceeding 99%. Watrous (1990) concluded that "acoustic phonetic speech recognition can be accomplished using connectionist networks" (p. 1753). Analyses of the trained networks indicated that the feature extraction units were trained to extract relevant spectral cues and not to abstract holistic templates. This shows that, in principle, the acoustic signal alone contains sufficient cues for phonemic discrimination with a single speaker and a single speaking rate and that these cues can be extracted automatically by statistical generalization of the sort an artificial neural network can be trained to do. This constrained case, however, poses a much easier problem than that of general word recognition in any speaker's natural running speech. The ad hoc nature of the networks' structure (a different one for each problem), and the neurally implausible training method, preclude more optimistic conclusions relevant to either human speech perception or machine word recognition.

The phenomenon of categorical perception, that is, the abrupt shift in stimulus labeling as acoustic properties are gradually manipulated and a corresponding relative difficulty in the discrimination of acoustically different stimuli that are perceptually assigned the same phonetic label, has received considerable attention since the earliest stages in speech perception research (see Repp, 1983, for a review). In a study of phonetic category acquisition, Seebach, Intrator, Lieberman, and Cooper (1994) used a biologically motivated network structure and training method to investigate whether categorical perception of stop consonants must be prewired in the brain or may be prenatally learned instead. Infants at a very young age are known to perceive categorically consonant–vowel (CV) syllables containing phonemes such as [b] and [p] (Bertoncini, Bijeljac-Babic, Blumstein, & Mehler, 1987; Eimas, Siqueland, Jusczyk, & Vigorito, 1971). The usual assumption is that the representations and processes that underly the categorization must be innate because there is insufficient time for infants to learn the categories. Infants, however, have functional hearing several weeks prior to birth, and it is possible that some phonetic categories are tuned prenatally on the basis of the impoverished auditory input available to the infant. Seebach et al. (1994) processed speech to simulate the intrauterine environment, and trained a five-"cell" BCM network (Bienenstock, Cooper, & Munro, 1982; Intrator & Cooper, 1992; see Figure 2) to distinguish between the syllables [pa], [ka], and [ta], spoken by a single speaker. After training, the units responded strongly to specific spectral characteristics of the input (e.g., low-frequency bursts), and the network successfully generalized (98%) to novel unvoiced-stop syllables from two different speakers (one male and one female) as well as to voiced-stop syllables (96%), even though no voiced stops were used in training.

The significance of this work with respect to the development of human speech perception is somewhat dubious, considering that neonates' phonetic categorization does not necessarily follow the boundaries of their native language (see reviews in Aslin, Jusczyk, & Pisoni, 1998; Jusczyck, 1997). For example, stop consonant categorization along voice onset time (VOT) continua, that is, as a

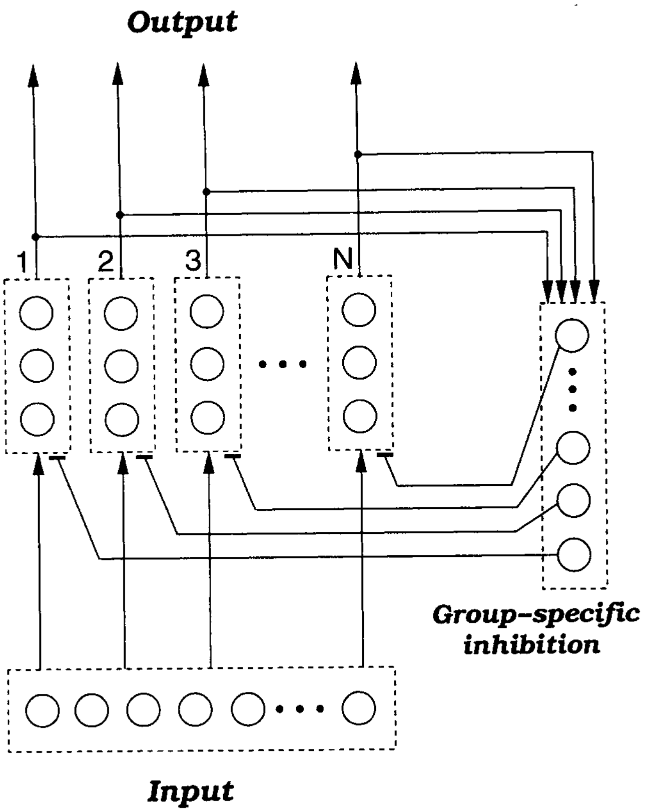

Figure 2. An example of a BCM network composed of *N* groups of 3 neurons each. All input features (nodes) are connected to all neurons in all groups. Summed output from all neurons in each group is fed into all neurons of the inhibition module, each neuron of which inhibits all neurons in a single group. The network learns through competition to inhibit all but the most active group, thus clustering the input patterns into discrete categories. From *Evidence for the Development of Phonetic Property Detectors in a Neural Net Without Innate Knowledge of Linguistic Structure,* by B. S. Seebach, 1990, unpublished doctoral dissertation, Brown University, Providence, RI. Copyright 1990 by B. S. Seebach. Adapted with permission.

function of the time between the burst from releasing the consonantal constriction and the onset of vocal fold vibration, is initially based on universal VOT boundaries and is later modified toward language-specific boundary values after considerable exposure to one's native language. It thus appears that initial perception of stop consonants is based on universal auditory features and not on spectral patterns learned in utero. In other words, there must be innate precursors to the phonetic feature detectors, which are not developed on the basis of prenatal auditory input. Whether such precursors are innate speech-specific feature detectors or more general auditory mechanisms remains to be investigated. The study of Seebach et al. (1994) only shows that, in principle, sufficient acoustic information is present even in degraded speech signals to enable the formation of appropriate phonetic categories.

Phones in Context: Learning Larger Prelexical Units

The simple CV syllables made up of an initial stop consonant followed by a vowel constitute an interesting special case of the general problem of phonetic identification because the acoustic realization of stop consonants is highly context dependent, so that this constrained problem is nontrivial, yet there are enough commonalities between the stimuli that it is not impractical to seek a relatively straightforward solution. Rossen (1989; see also Anderson, Rossen, Viscuso, & Sereno, 1990) used a neural network model to identify such syllables even in the presence of noise. Multiple two-dimensional intensity maps were used in the input representation as it was demonstrated that no single map could outperform the combination of all three maps together. The model consisted of a three-layer feedforward module trained with back-propagation (see Figure 1, right), followed by an autoassociative iterated map using the Brain-State-in-a-Box (BSB) algorithm of Anderson, Silverstein, Ritz, and Jones (1977). Output representation was localist, with a four-neuron group assigned to each of nine possible output phonemes (/p/, /t/, /k/, /b/, /d/, /g/, /a/, /i/, /u/). The network was trained to activate both the consonant and the vowel phoneme of each CV syllable presented, effectively acting as a syllable-recognition network. Apart from the argument for redundant representations, the contribution of this model must be evaluated with respect to its output representation. In particular, the syllabic nature of the network's responses is in agreement with recent arguments for the need (or the existence) of a syllabic level of representation in human speech perception (see reviews by Dupoux, 1993, and Eimas, 1997).

Principled arguments based on the university of syllables and the relationship between speech perception and speech production, as well as empirical arguments based on experimental findings supporting a syllabic representation, can be formulated to defend this position. The model of Rossen (1989) shows that multiple acoustic cues combine to yield fairly accurate estimates of the components of spoken syllables. Even though the output is expressed in phonemes and not unitary syllables, it is phonemes in context, as opposed to isolated phonemic segments, that the network identifies. It remains to be determined whether phonemes in context must be entire syllables as commonly understood in linguistics; that is, whether structural constraints on the possible orderings of phonemes are treated as describing entire syllabic frames (or templates) or, alternatively, simpler context-dependent phonemic units of representation are the appropriate level of description. Given the high degree of context dependence (coarticulation) found in natural speech, it is not surprising that context-dependent models have been proposed to deal with the intricacies of the multiple and sometimes conflicting cues to phonemes. It thus seems reasonable to attempt to model phonemic context dependency using linguistically meaningful units, such as syllables, for the representations.

In the same genre, but with the emphasis on a different issue, the three-layer feedforward network of Elman and Zipser (1988) was trained with the back-propagation algorithm to distinguish between CV syllables containing a voiced stop consonant ([b], [d], or [g]) and one of the vowels [i], [a], or [u]. Much like in the networks mentioned above, the input to this network consisted of temporally aligned spectrograms and was presented all at once. Training the network to assign the appropriate phonetic labels (consonant identity, vowel identity, or syllable identity) took more than 100,000 training cycles and resulted in an overall accuracy of about 95% (based on vowels and consonants). It was observed that distorting the input by adding noise to it was necessary to achieve the highest performance. The explanation for the advantage of using noisy input was that the idiosyncracies of particular training exemplars were blurred by noise, thus enabling the network to form better generalizations. This is an important issue in statistical learning where the specificity of training to the available input set must be offset by the need for successful generalization over category-equivalent test exemplars not present in the training set. Most interesting was the analysis of the representations developed at the hidden layer. Specifically, hidden units were found to maximally respond to particular kinds (groups) of sounds, such as vowel sounds or consonant sounds. It appeared, in other words, that over the course of learning the internal representation of the network developed in the way most appropriate for the task as it is typically conceived in theoretical treatments of speech.

Repeated simulations with different initial random weights gave rise to similar but not identical representations, leading Elman and Zipser to propose using multiple networks, each using a different internal representation. In simulations where the network was trained separately on the vocalic and on the consonantal portions of the syllables, hidden units were found to respond maximally to particular phonemes, such as an alveolar stop, in spite of the acoustic variability in the tokens of the phoneme in different contexts. On the basis of these findings, Elman and Zipser argued against strong nativist claims about speech representations. However, their claims must be tempered by the fact that the training method (including the learning algorithm, the nature of input representations, and the number of training cycles) does not seem to have much in common with biological learning, to the extent the latter is understood at all.

Mapping Auditory to Phonemic Space

Phonetic categorization implies a distorted mapping from input space to perceived (categorical) space, whereby large uniform regions in input feature space map onto smaller cluster regions in phonemic space that are maximally distinct. For example, stimuli differing in VOT in uniformly distributed steps map onto two distinct categories, within each of which discrimination between stimuli is very difficult. In contrast, discrimination between exemplars that belong to different categories, though not further from each other in feature (e.g., VOT) space, is very easy. This mapping is performed by the neural network models effortlessly by clustering input representations and associative, competitive, or supervised learning. In the case of vowels, a side effect of this distorted mapping is that vowels equally distinct in acoustic F_1–F_2 space (first and second formant frequency, i.e., frequency of the first and second energy peak in the acoustic spectrum) sound more similar when they are near the category centroid (in formant space) than when they are farther away from it. This effect, known as the *perceptual magnet effect,* can be observed as follows: for a pair of exemplars from a vowel category, discrimination between the two is more likely when they are close to the category centroid (prototype) than when they are farther away from it (but equally distant from each other; Kuhl, 1991; Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992).[1]

Guenther and Gjaja (1996) used a simple competitive neural network with formant frequency input and an auditory map composed of topographically organized nodes with fixed inhibitory connections to one another to demonstrate the formation of vowel category clusters in auditory space. The clusters were formed by unsupervised competitive learning after exposure to exemplars naturally distributed (i.e., not uniformly but normally around vowel centroids) in the formant space. No category labels or correct responses were ever presented to the network; the model self-organized to reflect the distribution of its input representation. The resulting auditory map closely modeled the perceptual magnet effect, including the formation of category-like clusters and prototype-like responses at the categories' centroids. It was thus shown that self-organization of a language-specific vowel space is possible given only formant frequency detectors and a neurobiologically plausible clustering processes through competitive interactions, without knowledge of target categories and, indeed, without prior knowledge that categories must be formed at all. As with the work of Seebach et al. (1994), the results of the simulations reported by Guenther and Gjaja (1996) must be taken into account, in that formant frequency detectors may need to be posited for language learning, but category-specific information and the fact that categories must be identified need not. The presence of finely tuned spectral peak detectors in the auditory systems of many animals, including amphibians, suggests that positing innate formant detectors in the human brain as well is not far-fetched. Again, it remains to be specified to what extent the formant frequency detectors in the human brain remain in the general auditory system or have evolved toward speech-specific functions.

Conclusion

In summary, the models reviewed above show that a good deal of phonetic information is present in the auditory signal and that mechanisms that extract this information can be found through the statistical generalizations of neural networks. However, with the exception of the model of Guenther and Gjaja (1996), the networks described offer no psychologically realistic options for human speech perception modeling. In addition, these networks lack compensatory mechanisms for several spectral properties with no linguistic relevance that stem from interspeaker differences, within-speaker variability, and from the transfer functions of the medium (e.g., recording equipment) and the surrounding environment (but see Carpenter and Govindarajan, 1993, for speaker vowel-space normalization using neural networks, and Grossberg, Boardman, & Cohen, 1997, for modeling speaking rate effects on phonetic categorization). On the other hand, auditory feature detectors are usually assumed to be innately present in the human brain. Networks, such as the ones presented here, may be trained under supervision to develop such detectors. The resulting processing structures would then preprocess the auditory signal, for a model of speech perception that could learn using pre-existing feature detectors like humans are thought to do.

Interactive Activation

The most prominent connectionist model in the area of speech perception has been, for a number of years, the TRACE model, as proposed by McClelland and Elman (1986; Elman & McClelland,

[1] But see also Aaltonen, Eerola, Hellström, Uusipaikka, and Land (1997), Lively and Pisoni (1997), and Sussman and Lauckner-Morano (1995) for more recent findings and interpretations.

1986) with some computational modifications (McClelland, 1991). Actually, there have been two distinct computational implementations: TRACE I was built to model data on phoneme perception, and TRACE II addresses issues of lexical access. The two versions share a number of architectural characteristics and mainly differ in their input representations and the stored lexicon. There are three functionally defined levels (or layers) of nodes: the feature level, the phoneme level, and the word level (see Figure 3). In a localist representational scheme, the TRACE model devotes one independent processing unit (node) to each representational unit in each level. In order to overcome the problem of temporal representation, each unit is repeated many times—once for each time slice. Each unit becomes activated when the units on the other levels that are consistent with it are activated. Activation is allowed to flow upwards as well as downwards, so word activation induces phoneme activation, which in turn may cause feature units to become activated (contingent on parameter settings). Naturally, feature activation causes phoneme nodes to become activated, which in turn activate word units. Special control parameters specify the computational details, such as the rate of decay and the type of influence between levels. For computational reasons, only positive (excitatory) activation is allowed between levels, whereas nodes inhibit other nodes within the same level. This way, nodes in one level indirectly inhibit nodes in the adjacent levels by exciting other nodes in them.

The model, illustrated in Figure 3, functions as follows: As time progresses, input is applied to successive time slices of feature detectors, but units at previous and upcoming time slices can be activated because of the overlap between features and phonemes and between phonemes and words. The final activation state of each time slice thus depends on large portions of the input, possibly the entire utterance, if more than one interpretation is locally possible. This way the model can also anticipate features because a word activated by its initial phonemes activates all of its phonemes, which, in turn, activate their corresponding feature detectors in future time slices before input to those time slices is present.

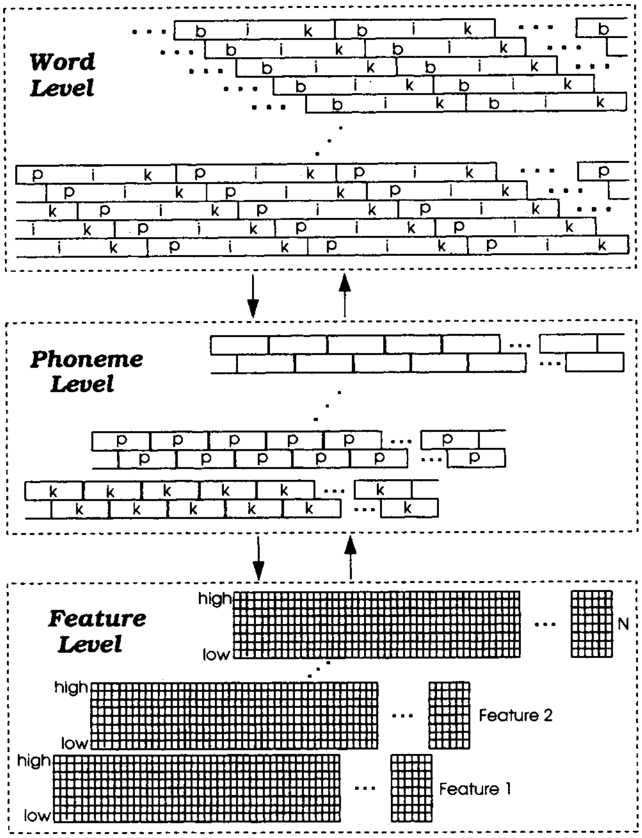

Figure 3. Schematic diagram of TRACE II, a model of phoneme and word perception, and of the interactions between the two (McClelland & Elman, 1986). As time unfolds, input is presented to successive units of the feature layer, from left to right, and activation spreads through the entire network (including past and future positions). At the lowest level, graded feature detectors are activated by acoustic analysis of the input at 5-ms intervals (in practice from idealized featural representations). Activation flows upward to phoneme units, each of which spans 11 feature units. Active phoneme units excite word units that contain them, send top–down activation to features consistent with their ideal composition, and inhibit other phonemes at the same temporal position. Active word nodes inhibit other word nodes occupying the same temporal position and send top–down activation to their constituent phoneme nodes. From "The TRACE Model of Speech Perception," by J. L. McClelland and J. L. Elman, 1986, *Cognitive Psychology, 18,* p. 9. Copyright 1986 by Academic Press. Adapted with permission.

Modeling Context Effects

TRACE I successfully handles contextual variability in acoustic–phonetic processing, including left-context and right-context effects and perceptual restoration of phonemes (i.e., differences in phonetic perception depending on preceding or following phonetic context and the inability to detect that a phonetic segment is missing when it has been replaced by noise, respectively). The context effects can bias interpretation of ambiguous stimuli or modulate the connection weights between particular features and phonemes depending on adjacent segments, thus potentially modeling phonological effects, articulatory constraints, as well as various other acoustic interactions between cues for different classes of phonemes. Simulations showed that weight modulation by activation of adjacent time slices improved voiced stop consonant identification from 79% correct to 90% correct, when presented followed by a single vowel (i.e., [b], [d], or [g] followed by [a], [i], or [u]). Weight modulation also facilitated identification of missing stop consonants (when the initial 175 ms were removed), a task in which human participants were found to perform similarly (Elman & McClelland, 1986).

TRACE accounts for trading relations found in the perception of stop consonants, such as the interaction between VOT and onset frequency of the first formant (F_1) (Summerfield & Haggard, 1977; see Repp, 1982, for a review and interpretation of the phenomenon). In a number of simulations, segments with lower F_1 onset needed higher VOT to activate /k/ as strongly as segments with higher F_1 onset. This continuity in stop consonant perception, in accordance with perceptual trading relations experiments, does not prevent TRACE from exhibiting categorical perception. The transition between voiced and unvoiced responses, as the feature specifications are moved along the VOT and F_1 continua, is much sharper than would be expected by the small feature differences because of competitive inhibition between the units at the phoneme level. The probability of correctly discriminating stimuli within a phoneme category is also much lower than the probability of discriminating stimuli in different categories, because activation feedback from the phoneme level to the feature level diminishes feature differences between patterns that correspond to the same phoneme.

TRACE II similarly models lexical effects, that is, differential processing of sublexical (including acoustic) representations depending on the lexical status of the result. For example, ambiguous segments are perceived unambiguously when in appropriate lexical context: A feature specification ambiguous between [b] and [p] activates /p/ more strongly than it activates /b/ if followed by [lʌg] because the word *plug* is thus formed, whereas there is no word *blug*. Ambiguous segments in single syllables are more likely to be interpreted so that the syllables follow the phonotactic rules of the language. For example, an input feature specification between [r] and [l] in the context of [s__i] activates /l/ much more strongly than /r/ because /sli/, although not a word, exists as part of words in the lexicon (*sleep, sleeve,* etc.), whereas /sri/ does not. Thus, there is no need for explicit specification of phonotactic rules because the statistical properties of the contents of the lexicon suffice to define them.

Because word activation builds up given matching input as time progresses, lexical effects in TRACE are predicted to be minimal on word-initial segments, rising gradually throughout the word, and much stronger on word-final segments; in the latter case, more activation has already accumulated at the lexical level to support compatible segments. This prediction has been challenged by results of subsequent phoneme monitoring experiments. Cutler, Mehler, Norris, and Seguí (1987) found lexical effects to be modulated by list composition, disappearing in monotonous lists, and appearing in more varied lists, and argued for an attentional mechanism and a serial dual-outlet model. Attentional modulation in noninteractive models was likewise proposed by Eimas, Marcovitz Hornstein, and Payton (1990) and Eimas and Nygaard (1992), on the basis of the interaction of lexical effects with word predictability and with secondary tasks. None of these effects can be accommodated in TRACE because there is no provision for stimulus-extrinsic factors, such as competing cognitive tasks, global attention, or long-term lexical statistics.

The nature of lexical effects has been a point of contention not only with respect to TRACE but more generally in speech perception research. In contrast to the predictions of TRACE, where such effects may occur at any point along a word, Frauenfelder, Seguí, and Dijkstra (1990) found lexical effects on phoneme identification only after the point at which a word becomes unique. They failed to find any lexically mediated inhibitory effects that would be predicted by the combination of top–down information flow and lateral interphoneme inhibition of TRACE. Moreover, McQueen (1991) found lexical effects on word-final phoneme identification only when using stimuli degraded by low-pass filtering, and then only in the faster response times (but see Pitt & Samuel, 1993, for a meta-analysis of the lexical effects on phoneme identification). According to TRACE, lexical effects should manifest themselves, if not independent of stimulus quality, certainly even with high quality stimuli. Finally, Pitt and Samuel (1995) found lexical effects early in the words and before the words' uniqueness points, as predicted by TRACE, but not gradually rising toward later positions in the words. Taken together, these findings suggest some limits to the interactive character of lexical-phonemic processing, at least of the type predicted by TRACE. The distinction between top–down expectation and activation of the adaptive resonance theory (reviewed below) may offer an alternate approach for a principled establishment of such limits. It appears more plausible that attentional mechanisms play an important role in constraining the flow of information along various paths and directions. It is at present unclear, though, how such mechanisms might be implemented in an interactive activation framework. Alternatively, simple recurrent networks, such as the dynamic-net model (Norris, 1990, 1992, also reviewed below), offer an approach to context effects with much less emphasis on top–down lexical information flow.

Word Activation and Recognition

With regard to the time course of word recognition, TRACE follows the basic guidelines of the early Cohort model (Marslen-Wilson & Tyler, 1980; Marslen-Wilson & Welsh, 1978), a fully interactive model of word recognition in which word units join a candidate list (cohort) as their onset specification matches the acoustic input and are then pruned when mismatches occur until a single candidate, fully compatible with the acoustic input, is left and thereby recognized. As in TRACE, later versions of Cohort (Marslen-Wilson, 1987; Marslen-Wilson, Brown, & Tyler, 1988) were modified to assign activation values, as opposed to a binary cohort membership-nonmembership distinction; in contrast to TRACE, the interactive character of the Cohort model was restricted to the initial phase of processing. In an additional departure from the Cohort model, TRACE can also handle ambiguous or distorted word onsets because words can be accessed at any point in their specification by compatible phonetic input, as long as they are correctly aligned. The relative importance of word onsets (Marslen-Wilson & Zwitserlood, 1989; cf. Connine, Blasko, & Titone, 1993) is also preserved, in that word-initial information is more effective in activating a lexical entry. This is modeled indirectly in TRACE and does not depend on any special assumptions about word onsets. Words activated by their initial segments contribute to the activation of their constituent phonemes by providing positive feedback, and words whose initial segments are not matched are suppressed by within-level inhibition and require stronger activation later to stand out as likely candidates. Recently, Allopenna, Magnuson, and Tanenhaus (1998) provided evidence for the weak rhyme effects that are predicted by this account (but precluded in Cohort) using an eye-tracking paradigm.

One robust finding in human speech perception is that the frequency of occurrence of a lexical item affects its probability of activation such that more frequent items are recognized faster than less frequent items and, in cases of phonetic ambiguity or impoverished acoustic information, are more likely to be recognized (Marslen-Wilson, 1987; Zwitserlood, 1989; see Bard & Shillcock, 1993, for review and a related discussion). Such frequency effects, although not currently accounted for by TRACE, might be accommodated by adjusting the resting activation of words. In models that incorporate a word learning stage with variable frequency of presentation between items such effects are naturally accounted for as a part of the models' statistical properties.

The problem of word segmentation is not addressed separately in TRACE because the speech stream is automatically segmented as a by-product of lexical activation. Therefore, there is no need for word-boundary cues or access-initiation strategies. As in human speech perception, in many cases a word can be successfully recognized in TRACE only after part of the next word is heard or even later (Bard, Shillcock, & Altmann, 1988; Grosjean, 1985; Luce, 1986). Finally, because of the temporal spread of the feature detectors and their connections to the phoneme units, the model is able to cope with assimilation phenomena at word boundaries. In the case that more than one word is possible, TRACE can come up with all possible alternative candidates for a decision to be made by interactions at higher levels of processing.

Criticisms of TRACE

According to McClelland and Elman (1986), the success of TRACE is attributed to its massively parallel, interactive processing and to its architecture, which directly implements the assumption of an utterance being decomposable into sequences of units at several processing levels. More recently, McClelland (1991) effectively reduced the entire model to an argument for interactive activation, in accordance with some researchers' intuitions about the processing of the brain. It is not clear yet that the results achieved by TRACE cannot be successfully modeled by more linear approaches, or that the type of processing that produces these results is similar to human speech processing. Massaro (1988, 1989) and Norris (1982, 1992, 1993), among others, have argued against interactive processing, putting forward noninteractive models to account for the same problems. Norris (1993), in particular, argued that "behavior that looks like interaction between processes can arise from learning, rather than from any genuine on-line interaction between processes" (p. 215). This position is discussed in more detail below.

A number of significant deficiencies of TRACE have been pointed out by McClelland and Elman (1986), mostly concerning the representation of time. Specifically, the requirement that each unit on each level be repeated for each time slice leads to a very inelegant representation with no known homologue in biological systems. This temporal slicing has very important implications for learning because whenever a new word is learned it must be copied to each time slice individually. In addition, each new word, presumably represented by a new node at each time slice, must be connected via bidirectional inhibitory projections to all other words, an unlikely feat for learning systems (cf. Grossberg, 1986). Another problem is that TRACE is presently unable to handle speaking rate variability, as well as other global differences such as speaker characteristics and accent. The failure of TRACE to deal with real speech is also a major shortcoming. The main reason for the inability of TRACE to handle real acoustic input may be the temporal variability of speaking rate (Elman, personal communication), a problem that had also baffled speech recognition engineers until the relatively recent emergence of dynamic programming.

Besides the fact that TRACE does not recognize speech, and is thus more of a tool for psychologists than an application for engineers, another limitation of TRACE is that it does not learn anything. It is prewired to achieve all its remarkable results, thus effectively encoding the knowledge and intuition of its designers. This is desirable in a way, insofar as there is a commitment to the perceptual mechanism. That is, the performance of the system does not rely on some intractable generalization of a powerful mapping network but, rather, reflects the researchers' knowledge about human speech processing. However, many aspects of human speech perception (e.g., the language-specific segmental specifications and contrastive feature sets) are learned during development, and, certainly, the lexicon is constantly enriched and expanded. Although the form of lexical representations and the way they arise (as a result of exposure to a linguistic environment) might be universal and genetically encoded, the phonological and phonetic structures of any given language are not. Therefore, a complete model should be able to derive the representations humans learn during development by exposure to the language, abstracting the appropriate regularities using some initial set of constraints. Furthermore, the hard-wired functional structure of TRACE and other Interactive Activation models may in fact be neurobiologically unlearnable (and not merely unlearned in the models' implementations), casting doubt on the psychological relevance of the simulations (Grossberg, 1987).

Other shortcomings of TRACE include the small set size in each of its levels (incomplete sets of phonemes and features and a small vocabulary), which might have significant consequences for the resulting behavior. It is unclear how the model would behave with a realistic lexicon, a full set of phonemes, and complete feature descriptions. McClelland and Elman (1986) only used a restricted set of phonemes and words to demonstrate particular effects, and in some simulations they even left out parts of the system that they considered irrelevant to the phenomena being modeled. Finally, it may be disturbing to some that a model of parallel distributed processing is, in fact, using localist instead of distributed representations, therefore missing out on an important innovation the connectionist field has brought about. In all, TRACE is certainly showing its age, and more modern approaches are needed to address the host of recent findings in the field. Indeed, alternative approaches have been developed that address many important issues not adequately covered by TRACE. The following sections are organized by reference to the main weaknesses of TRACE, namely, the representation of time, the need for learning, and the role of phonological structure in the lexicon. Each of these issues is first introduced generally and approaches to resolve it are then discussed.

Temporal Representation and Integration

A problem that comes up in speech modeling and in other language-related domains is that they are intrinsically temporal. Speech, in particular, is represented in an acoustic waveform, which is air pressure (or spectral energy) as a function of time. To represent that in a neural network, most earlier models would just present to the network an entire time frame simultaneously, with a unit or a subset of units devoted to each of a number of time slices, possibly differentially weighting the slices by distance from the current time slice. This way, one avoids rather than solves the problem of time, perhaps in a way very unlike the brain (cf. the naive view of time in Port, Cummins, & McAuley, 1995).

Spatializing Time

Direct time-to-space transformation can be thought of as a special case of time-delay architecture, whereby special nodes in the network represent input at certain time differences. Time Delay Neural Networks (TDNNs) are usually realized with delay units between adjacent nodes and have been proven useful in speech recognition (Haffner & Waibel, 1992; Lippman, 1989; Waibel, Hanazawa, Hinton, Shikano, & Lang, 1989). Waibel (1989) examined internal representations formed by a TDNN trained to discriminate between the three voiced stop consonants ([b], [d], and [g]) and found them encoding linguistically plausible features such as formant movement detectors and boundary detectors. In addition, hidden units were found to operate in a (temporally) shift-invariant manner, leading to phoneme recognition without explicit segmentation. More recently, Windheuser and Bimbot (1993) trained a TDNN to recognize distributed phonemic representations, in terms of binary phonetic features, as opposed to the usual single-unit representation per phoneme. Using real speech input from a single speaker, the network achieved its highest performance (95.9%) using an expanded, redundant feature set and more detailed labeling of diphthongs.

Using a neural network with time-delayed temporal representations for speech processing is equivalent to making a critical assumption about acoustic representation, that is, that there is a short-term storage of the exact acoustic information from previous times in an unprocessed form. Although not impossible, it is rather unlikely that the brain functions this way; temporal axes have so far been identified only in sound localization structures in noncortical nuclei. Rather, it might be better to include the temporal information in the network structure, so that the context information is instantiated as a modulation of the processing properties. We have been unable, so far, to identify storage structures in the brain that are not processing structures as well. One of the reasons that neural networks have such an appeal for cognitive modeling is that storage and processing in neural networks are integrated and indistinguishable in the connection weights and the spreading of activation, unlike in digital computers and other symbolic machines where there is a processor that acts on data that are stored separately from the program. By using TDNNs, one reduces the networks to static statistical approximators, thus abandoning one of the primary reasons for using neural networks in the first place, that is, to use brain-like processing of brain-like representations.

Fortunately, alternative architectures have already been developed for representing temporal sequences in neural networks (see review in Mozer, 1993) and, more specifically, for speech-related simulations (Elman & Zipser, 1988; Lippman, 1989; Norris, 1990, 1992). However, it is still more difficult to train some temporally varying models than static models like TDNNs. The known training methods require long training times, large amounts of computer space, and tend to be unstable as well. A popular option is back propagation of error through time, which is essentially an unfolding of the network by duplication for each time step, followed by regular back propagation of the error from an arbitrarily distant point in time (Pearlmutter, 1989). A more economical alternative is to include delays for the recurrent units and train the network by copying the previous values to the next time step and then using standard backward error propagation (Elman, 1990; Norris, 1990).

The Elman-Norris Net

Previous work by Jordan (1986) on sequence production had demonstrated that adding delayed recurrent links from a network's output units back to the hidden layer through a *state layer* (see Figure 4, left) enabled the network to encode and reproduce sequences of vectors. The addition of the state layer acted as a previous-context layer, feeding back into the hidden units the network's prior output states in parallel with the new input. Connections from the state units back to themselves ensured encoding of sequences of arbitrary length because the prior context was also influenced by its own prior context. An additional advantage of this type of recurrent network is that the recurrent links may have fixed weights because they only need relay prior state and all computations can be performed by the forward connections. Thus, the standard back-propagation algorithm can be used for training, avoiding the computational expense and other complications of other recurrent training methods.

Elman (1990) modified Jordan's architecture to model temporal pattern classification and series prediction. In contrast to Jordan's net, which needed to compute the next item to be *produced* in a sequence based on the currently active *output* item, a temporal pattern *classification* network (like a speech recognition network) must identify a given *input* sequence based on the current plus prior *inputs* and classify each sequence appropriately (cf. Norris, 1990). Thus, the recurrent links were connected from the hidden layer and not from the output layer, to store previous input context and, recursively, the context of previous context and so on, because there were closed loops in the connections. As already mentioned, such a network can be trained using standard back propagation. Training to predict the next item in a sequence, as opposed to a variant of the auto-association task, forces the network to develop representations that model the temporal structure of the input pattern sequences. Simulations showed that the network accurately predicted future segments when the regularities in the training sequences provided the appropriate information (Elman, 1990). Because the prediction error was maximal between learned sequences, and very small within such sequences, Elman (1990) also proposed that a network of this type may be used to discover word units in an input stream by hypothesizing boundaries at points of low predictability.

Treating speech as a sequence of events, rather than as a spatialized pattern of activation, Norris (1990) proposed a *dynamic-net model* that also included recurrent connections within the hidden layer (Figure 4, middle). The network was trained to identify words (treating phoneme outputs as "don't care" conditions) and phonemes (likewise ignoring word outputs) in alternating training epochs. Early versions of this network showed insensitivity to changes in input rate and shift-invariant sequence recognition, as desired (Norris, 1990). More recently, lexical effects on phoneme identification, phoneme restoration, and compensation for coarticulation were exhibited by networks of this kind (Norris, 1992). It was thus shown that a recurrent network can perform word recognition by preserving context in its hidden layer, without resorting to artificially spatial representations of time (like that of TRACE).

On the basis of the performance of this kind of recurrent networks, Norris (1992) made a number of interesting remarks on the nature of interactive and top–down processes. He argued that

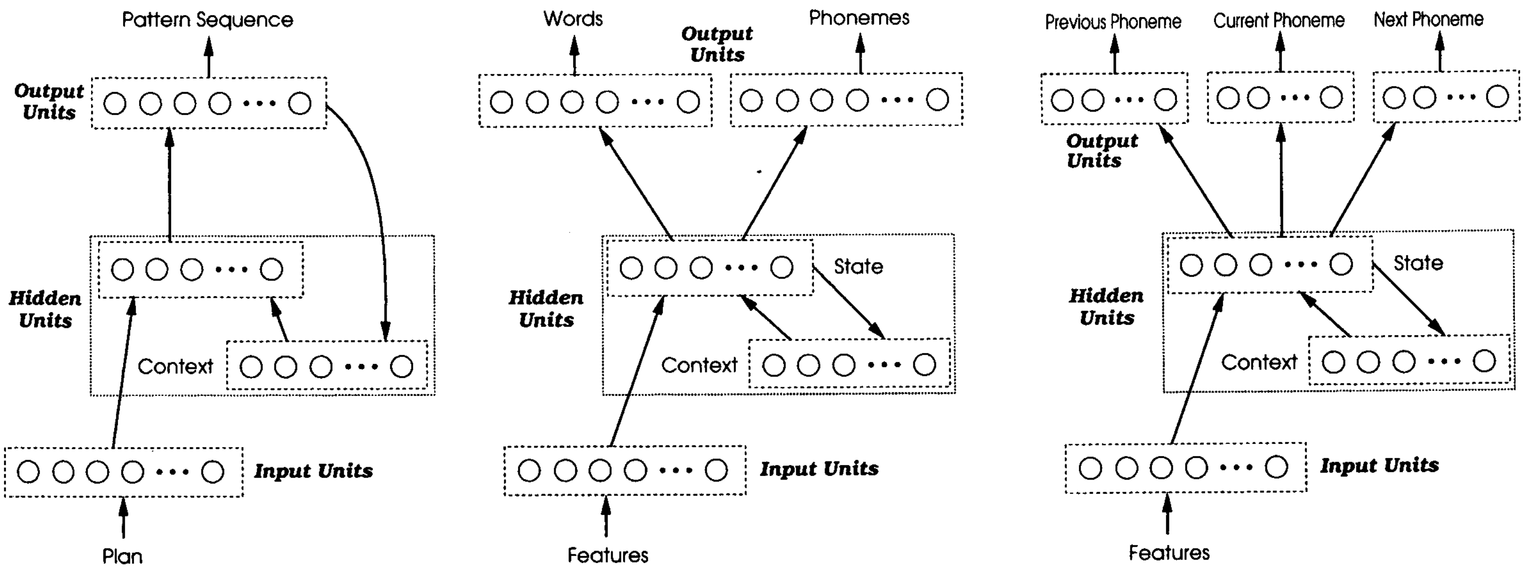

Figure 4. Simple recurrent networks with a context layer that presents the state of the network along with new input. Left: Network architecture proposed by Jordan (1986) to associate static input plan vectors with entire sequences of output patterns. The recurrent connections allow the network to know its previous output state to modify subsequent behavior appropriately. From "Finding Structure in Time," by J. L. Elman, 1990, *Cognitive Science, 14,* p. 183. Copyright 1990 by the Cognitive Science Society. Adapted with permission. Middle: A dynamic-net model of human speech recognition (after Norris, 1990, 1992). Phoneme and word outputs are independent of each other and are trained separately, but the same hidden units are used for both. During testing, input feature sequences activate both the appropriate current-phoneme output node and the appropriate current-word node. Note that, although the topology of the network layers and connections follows Norris (1992), the functional characteristics implied here by the terms *context* and *state* (cf. Gaskell, 1994; Gaskell et al., 1995) may not coincide with Norris' views (see text for details). From "Finding Structure in Time," by J. L. Elman, 1990, *Cognitive Science, 14,* p. 184. Copyright 1990 by the Cognitive Science Society. Adapted with permission. Right: The Elman net used by Shillcock et al. (1991) and Shillcock et al. (1992). The phonetic features of the input plus the context that is saved in the recurrent hidden layer are mapped onto a featural specification of the prior, the current, and the predicted phoneme. The featural specifications are identical for the input and the output node groups but differ in the two implementations by Shillcock et al. (1991, 1992). From "A Connectionist Model of Auditory Word Perception in Continuous Speech," by R. Shillcock, J. Levy, and N. Chater, 1991, in *Program of the Annual Conference of the Cognitive Science Society, Vol. 13* (p. 343), Hillsdale, NJ: Erlbaum. Copyright 1991 by the Cognitive Science Society. Adapted with permission.

the two are distinct concepts and that it is sometimes very difficult, if at all meaningful, to attempt to classify connectionist models as such. He also argued that backward error propagation during learning is a form of top–down information flow, although during testing activation is only allowed to proceed in one direction. In Norris' view, "talk of interaction implies that we have identified two or more processes in the model that might potentially interact," whereas "if we want to classify a model as 'top–down' or 'bottom–up,' all we need to know is the direction of information flow as it passes from input to output" (1992, p. 367). In a network like the dynamic-net, it is not possible "to slice the model up into two discrete stages corresponding to word and phoneme recognition" because, with the exception of the weights between the hidden units and the output units, "everything else in the network is involved in computing some intermediate representation that subserves both word and phoneme recognition" (Norris, 1992, p. 368). Norris concluded that his model is entirely bottom–up, claiming that information flows in one direction only, and argued against the need for interactive models of speech processing such as TRACE.

An alternative interpretation might be to consider an effect top–down in phonetic processing, if lexical information of any sort can influence phonetic decisions. This would imply a theoretical commitment to treating phonetic processing as essentially lower level than lexical processing, even though the simulations reported by Norris (1992) indicate that in practice such segregation is not necessary. Because the hidden layer in the dynamic-net model encodes both lexical and phonemic sequential information (in other words, the hidden layer encodes feature bundles in context), the model's decisions about phonemic identity cannot be assumed to be free of lexical influences. However, the view that there is no top–down flow of information may be an artifact of the way the model is schematized. In Figure 4 (middle), the hidden layer has been drawn in a manner functionally equivalent to having within-layer delayed connections, but insofar as previous context is considered input to the hidden layer, parallel to current input, the recurrent connections may arguably be considered top–down information flow. For this reason, the computing hidden units here have been named state units because they encode the present state on the basis of which the output layers indicate the presence of a phoneme or word in the input. The relay hidden units have been named context units because they encode the prior information in the context of which the computations of the input are performed.

Shillcock, Levy, and Chater (1991) used a similar recurrent network (Figure 4, right) with three output node groups corresponding to the previous, current, and upcoming phoneme, all trained simultaneously, to model the phoneme restoration effect and data on phoneme monitoring (see also Levy, Shillcock, & Chater, 1991). Their model, too, showed phoneme restoration for slightly degraded input, without any reference to word frequency or similarity. In fact, the authors claimed that this kind of network captures the phenomenon more accurately than TRACE because in the latter word activation may override acoustic information, effectively hallucinating phonemes. Shillcock et al. (1991) used a binary featural representation, on the basis of the acoustic features of Jakobson, Fant, and Halle (1972), for the model's input, and phonemes for its output layer. Training of the network was done using the copyback technique. An extended output window was used to illustrate expectancies and right context restoration. A more recent version (Shillcock, Lindsey, Levy, & Chater, 1992) used a featural input representation motivated by current developments in phonological theory, again using mostly binary values. Most interestingly, context effects were successfully modeled even though there were no localist lexical representations in the model, allowing Shillcock et al. (1992) to argue for a direct mapping of the (phonetic) featural representation to the semantic level, that is, a distributed representation of the lexicon.

Gaskell and Marslen-Wilson (1997) also proposed a similar model with a distributed lexical representation, that is, in the word-output layer (of a network like that in Figure 4, middle), each node did not correspond to a single word but participated in the representation of many words. The implications of modifying Norris' net to use a distributed output representation are most important with respect to the concepts of activation and competition because, in contrast to a model of localist lexical output, there is no direct correspondence between node activation in a distributed representation and lexical item activation. Rather, the distance between the layer's total activation pattern and each word's representation is taken to quantify word activation. In this way, simultaneous activation of many lexical items is realized by an activation pattern in the model that is intermediate between the activation patterns corresponding to the lexical items involved. Competition between lexical items cannot thus be encoded as activation flow between nodes and its behavioral consequences can only be evidenced indirectly in the way word "activation" (i.e., distance between patterns of node activation and of lexical specification) is affected in the model's function as processing time progresses. Notably, the network does not rely too much on similarity between items, which would cause it to settle for intermediate (nonword) activation patterns, but displays the desired sensitivity to input phonetic features that has been shown in human speech perception. It remains to be demonstrated that this kind of distributed representation is capable of modeling the complex behaviors successfully captured by TRACE and the dynamic net, and that the observed advantages are indeed a consequence of the representational scheme and not of the parameter setting of the particular model.

The Shortlist Model

The debate notwithstanding over whether such a context-preserving model may be called strictly bottom–up and what exactly that would mean, the dynamic-net model shows that a recurrent neural network can be successfully trained to recognize words from sequences of feature bundles using an internal representation of time in the form of context-encoding units. To the extent that one can account for experimental findings usually cited as evidence to support TRACE using this architecture, the dynamic-net model constitutes a considerable improvement. It is not the case, however, that such a simple net alone can incorporate all the functionality of TRACE, in particular with respect to right-context effects and the selection of words among competing candidates. For example, given the input [kɒrpə], the network cannot know whether *carpet, carpenter,* or *car pollution* is the correct parsing before receiving information about several more segments. As already discussed, TRACE takes care of this problem by duplicating the lexical network for each possible temporal alignment and by competition (through mutual inhibition) between active words occupying the same temporal position. To address this issue, Norris (1994) added a competition network, on top of the recurrent recognition network, in which words detected in the input stream were entered as candidates and allowed to compete with each other, depending on their phonological overlap. The competition network is wired on the fly by a program from the list of most active word nodes of the word recognition network. Because only a few words are active enough to be used in the list, the model was named Shortlist (Norris, 1994).[2]

The operation of Shortlist is as follows: Ideally, a dynamic-net model first takes featural input from analysis of the speech signal and activates output nodes corresponding to words that contain these feature-bundle sequences. In practice, phoneme strings were looked up serially in an electronic dictionary to make implementation of the final stage possible using a realistic vocabulary, without the need to train a huge word-recognition network. At each time step, the most active word nodes are entered into a shortlist, and their phonological constituents are looked up to determine the degree of overlap between them. A small competitive network is then wired (see Figure 5a), in which the words are allowed to inhibit each other for a fixed number of iterations before their resulting activations are noted. The entire process is repeated with the next phoneme or, in the full (unimplemented) model, with the next bunch of features. In fact, the competition network is an interactive activation network equivalent to the word level of TRACE, but with all subthreshold words and connections to them removed. The proportion of connections and nodes thus removed is very large and results in tremendous savings on computations, making a large-vocabulary implementation possible, whereas the result is not affected because only inactive (or almost inactive) nodes and connections are ignored.

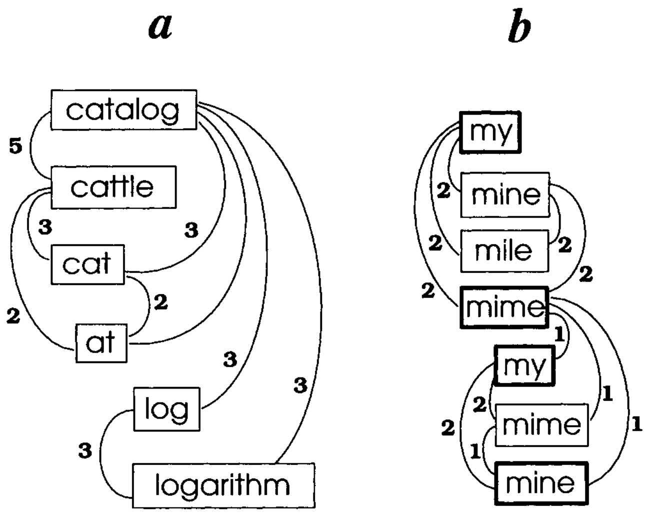

Figure 5. Examples of Shortlist competition networks. (a) Competition network after presentation of the string [kætəlag], only including candidates that perfectly match the input (not all candidates are shown). Note that, as candidates are entered into competition after discrete time steps in the lexical search, time-alignment information is available to align the word nodes appropriately, resulting in a network identical to the word level of TRACE with all subthreshold nodes, and the connections to them, removed. Between-word inhibition is proportional to the degree of overlap, indicated here by numbers (of phonemes) next to the inhibitory connections. (b) Competition network after presentation of the string [maɪmaɪn], including candidates matching the input to various degrees (not all candidates are shown). Candidates perfectly matching the input are shown in bold outline. Note that the alternative interpretations may only be resolved after subsequent context is presented. In the meantime, multiple activation of the same word is possible without confusion because competing nodes are unrelated tokens. In addition, the temporal alignment allows even tokens of the same word to compete appropriately with each other (e.g., here two tokens of *mime* compete for a common /m/).

Shortlist successfully simulates activation of lexical items that match the phonemic input, competition between them, and selection of the item combination best allocating all input phonemes. McQueen, Norris, and Cutler (1994) recently tested the predictions of Shortlist regarding inter-word competition using a word spotting task. They found that words embedded in word beginnings (such as *mess* in [dəmɛs], the onset of *domestic*) are harder to spot than the same words embedded in phonemic strings that don't begin words (e.g., *mess* in [nəmɛs]). In addition, imperfect phonemic matches (using input that is ambiguous, or that deviates from the internal specification by a few features) activate the appropriate lexical candidates if the mismatch is not too large (e.g., [ʃɪgərɨt] activates *cigarette*), and longer words are more resistant to degradation than shorter words.

In all, the dynamic-net model offers an attractive alternative to TRACE with respect to temporal representation, and Shortlist augments the system with a necessary competitive stage, whereby candidate lexical items are selected on the basis of both prior and subsequent context. The activation levels of the word nodes in the small competitive network seem to model well findings from the cross-modal priming task regarding the time course of multiple word activation, competition, and selection (e.g., McQueen et al., 1994; Norris, McQueen, & Cutler, 1995). Furthermore, although phonemic processing is certainly affected by lexical information because of the shared hidden layer and its recurrent connections, word activation that results not from bottom–up information flow, but from interlexical competition, does not influence phonemic processing. This suggests that there is a very specific (and testable) limit to the possible lexical effects on phonemic categorization and their time course. More specific simulations are needed to investigate the relationship between word length and word overlap (human data reviewed and TRACE simulations reported in Frauenfelder & Peeters, 1990), neighborhood effects (how the number of phonemic *neighbors,* i.e., lexical items differing by a single phoneme, affects lexical processing; Cluff & Luce, 1990; Goldinger, Luce, & Pisoni, 1989; Luce, Pisoni, & Goldinger, 1990), and the as yet inconclusive findings on word-final embedded priming (such as activation of the word *bone* on hearing *trombone* or the word *lip* when *tulip* is heard; cf. Gow & Gordon, 1995; Shillcock, 1990; Tabossi, Burani, & Scott, 1995).

With respect to the remaining shortcomings of TRACE, namely, lack of a learning mechanism and implausible architecture, Shortlist constitutes little improvement. Specifically, back propagation

[2] The list is also kept short, in the case of many active candidates, by limiting its maximum length to cut down on computational cost. Simulations showed that a limit as low as two allowed the model to function appropriately. A limit of 30 was used in the simulations reported by Norris (1994).

is no less biologically implausible than hard-wired connections (Grossberg, 1987, pp. 47–50; see O'Reilly, 1998, for a biologically plausible version of error-driven learning). The plausibility of on-the-fly setting up of a whole new competitive network and its connections at every time step is also questionable. The fact that a given vocabulary may be learned during training is certainly an improvement, but how the learned phonemic and lexical information of the recurrent network is used to align the candidate items in the shortlist is left somewhat unspecified at this point. In particular, it is still unclear how the lexical and phonemic nodes are combined to yield the phonemic representation of each lexical item that is necessary for setting up a competitive network. Furthermore, the network is certainly unable to learn more new words once trained without either forgetting others already learned or needing representation of the entire training set. The model's treatment of nonwords (or novel words) is also unclear. Finally, Shortlist offers no improvement over TRACE with respect to its input, which is also far removed from a realistic speech signal. On the other hand, Shortlist is a relatively recent model compared with TRACE, and future investigations will likely address many of these issues and other possible shortcomings.

Learning and Development

Most connectionist models, with the notable exception of TRACE, are trained on a set of data to generalize to another set of data. This is usually accomplished by successive presentations of an activation vector to the input layer of the network, and comparisons of the network's actual output to the desired outcome, followed by some sort of weight adjustment to reduce the observed discrepancy. Alternatively, connection weights may be tuned through competition and reinforcement on the basis of internal activation patterns only (i.e., without explicit error computation). For example, correlated inputs can cluster into categories by activating the same units and strengthening connections between these active units. Active nodes may also compete for activation so that prototypical input patterns end up strongly activating a single output node.

For the purpose of building a functional model of speech perception, it is not necessary to include a learning component (TRACE is a case in point). Nevertheless, to construct a complete model, plausible not only biologically, but also developmentally, the issue of network modification as a result of interaction with the environment must be addressed. In addition, models that are trained by a neurally plausible form of learning, such as BCM (Bienenstock et al., 1982) and ART networks (discussed below), may be more likely to develop representations similar to those of the brain.

Adaptive Resonance Theory

The Adaptive Resonance Theory (ART), proposed and continuously refined through the years by Grossberg and colleagues at Boston University, combines neurobiological plausibility with mathematical rigor to account for a host of psychological and neural findings, including memory, learning, attention, priming, pattern recognition, etc. (a recent review can be found in Carpenter & Grossberg, 1995). A detailed description of even one of the many versions of ART networks that have appeared in the literature is well beyond the scope of the present review; the reader is referred to Carpenter and Grossberg (1991) for a more complete discussion of ART and its applications. A reading of Grossberg (1986) is recommended for a palatable yet rigorous introduction to the ART formalization and concepts, with emphasis on speech-related issues.

In this section, the ART is briefly sketched, with emphasis on its speech perception modeling applications, because it encompasses a biologically plausible and psychologically relevant learning component.[3] Learning is incorporated in ART as an intrinsic part of the functional system and not, as in most neural network models, as an initial preparatory phase of statistical generalization. ART networks learn continuously, although the learning rate and the conditions that must be met are adjustable and may be influenced by attentional demands as well as by age constraints. Furthermore, learning in ART networks does not involve any separate, biologically implausible, mechanism: Weights are updated according to a differential equation as are the units' activations, only at a slower rate, on the basis of the correlation of activation flow with postsynaptic activation. Multiple ART modules may be combined to associate categories at different levels of representation, as well as between lists and their component units (Grossberg & Stone, 1986).

An example of an ART network is illustrated in Figure 6, including (a) an input layer I that registers an incoming vector (or vector sequence in the case of dynamic input), (b) a feature layer F_1 that admits bottom–up input from I and top–down expectations from F_2, and (c) a category layer F_2 that learns stable categories, on the basis of the featural patterns activated over F_1, and projects expectations based on its own pattern of activation back to F_1. The connection weight matrices between the two layers, called *adaptive filters,* are learned on the basis of the network's experience and adjust the flow of activation between the layers. ART networks are different from most networks of other architectures on several key points, including combining stable learned categories and the ability to learn new ones, top–down template learning without weight transport, matching between top–down expectations and bottom–up patterns, and continuous associative learning integrated in the system's function (Grossberg, 1987).

An implicit conceptualization in many connectionist systems, the distinction between short-term memory (STM) being an activation pattern and long-term memory (LTM) being a connection weight matrix is rarely made explicitly and in this terminology, as it is in ART. The two kinds of memory are treated in a unified manner, in terms of differential equations expressing the connection weight values and the activation of each node as functions of time, connection weights, and node activation, albeit in different time scales; node activation and decay are much more rapid than alterations of LTM traces. In contrast to most other existing connectionist systems, ART learning is performed noncatastrophi-

[3] By *psychologically relevant,* it is implied that the model's learning conditions and states can be cast in terms of psychological significance. For example, short-term memory activity, realized as an adaptive resonance, leads to modification of a category prototype, whereas failure to recognize an exemplar affects vigilance and attentional gains that lead to formation of new categories by strengthening connections to uncommitted patterns (Grossberg & Stone, 1986).

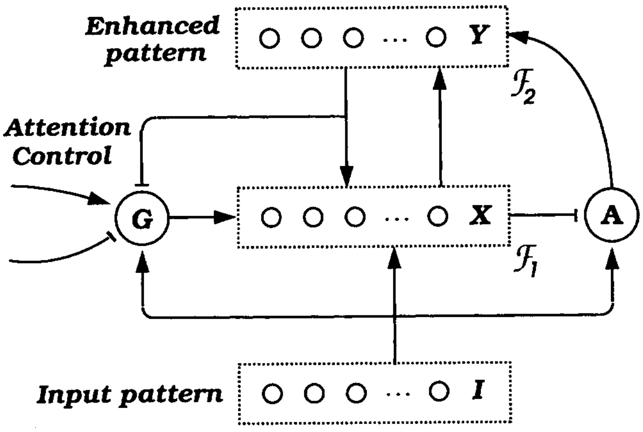

Figure 6. Basic architecture of an adaptive resonance module (based on Grossberg, 1987, and Grossberg & Stone, 1986). A preprocessed input pattern I is registered at the feature detector layer F_1, giving rise to a short-term memory (STM) activation pattern X. The existence of input activation I also excites the attentional gain control G and the orienting subsystem A, but the latter is inhibited by the STM pattern X in F_1 and does not become activated. The connection weights, called long-term memory (LTM) traces, between F_1 and F_2 form an adaptive filter, through which activation flows upward to the coding level F_2. As a result, a higher-level STM pattern Y is formed in F_2, which is enhanced by interactions between the F_2 nodes. F_2 activity inhibits the attentional gain control G and projects a learned top–down template (expectation) down to F_1 through a distinct adaptive filter. Because of the F_1 nodes' firing threshold and the lack of attentional input from G, the resulting overall X^* activation at F_1 will be lowered if the top–down expectance mismatches the bottom–up STM pattern, thus releasing inhibition of the orienting subsystem A. Activation of A nonspecifically resets active F_2 nodes to allow a new STM pattern to emerge and another top–down expectation to be matched against the STM pattern in F_1, until either a sufficient match is formed or the stored templates are exhausted. In the case of a good match between top–down and bottom–up pattern activation in F_1, X^* and Y excite each other and result in a system resonance that lasts long enough for the slow-changing LTM traces to be updated. External nonspecific attentional signals through G may override bottom–up information in particular circumstances. From "Competitive Learning: From Interactive Activation to Adaptive Resonance," by S. Grossberg, 1987, *Cognitive Science, 11,* p. 35. Copyright 1987 by the Cognitive Science Society. Adapted with permission.

cally for the existing categories that have been previously formed in the system. Input patterns deviating substantially from all stored categories do not destroy the existing structures because of the combined efforts of an attentional gating system that matches learned expectations to sensory patterns and an orienting subsystem that resets active categories that fail to produce an acceptable match to the input (see Grossberg & Stone, 1986, and Grossberg, 1987, for more details on the system's operation).

A property of ART networks that is very important for speech perception modeling is the distinction between top–down and bottom–up activation that makes it possible to prime STM activation of expected patterns but not actually activate them without sensory input. Referring to Figure 6, the attentional gain control G is constructed to output a nonspecific excitatory signal into the sensory feature level F_1 when excited by some attentional gain mechanism. The attentional gain may be activated by factors outside the speech perception system (e.g., task-specific attentional allocation) or by speech input. Because of their activation threshold, nodes in F_1 will not generate output unless excited by at least two of the three possible sources (I, G, and F_2). If the input matches the top-down "expected pattern" then F_1 nodes remain suprathreshold, but if there is a mismatch, then individual F_1 nodes will only receive input from one source, either from I or from F_2, and will be forced to shut off. Importantly, top-down active templates alone cannot give rise to F_1 pattern activation unless a matching sensory input is registered. The conceptual parallels of this analysis to the well-known priming effects in spoken word recognition are obvious: Associates of heard or seen words will be active in F_2 but not in F_1 before sensory input in their support can be registered. Interactive activation effects (often modeled with reference to TRACE) can be accommodated without the embarrassing tendency of TRACE to hallucinate phonemes because of word-node activation without bottom–up support.[4]

ART networks are capable of learning new categories without forgetting existing ones and without the need for an external omniscient teacher. In the case of a good match between top–down and bottom–up pattern activation in F_1, X^* and Y (see Figure 6) excite each other and result in a system resonance that lasts long enough for the slow-changing LTM traces to be updated, enhancing the category's representation. The orienting subsystem A serves to indicate when an acceptable match or a substantial mismatch has occurred. The capability of ART networks for constant learning from specific new exemplars may account for speaker-specific effects on word recognition (Mullenix, Pisoni, & Martin, 1989; Nygaard, Sommers, & Pisoni, 1994; see reviews and discussions in Johnson & Mullenix, 1997). The LTM traces do not reflect an abstract representation of words, but an average of all prior occurrences of a word with the latest speakers' utterances least decayed (recency effect). Therefore, words spoken by either a very well known or a recently heard speaker will match the stored patterns best. This way one can also adapt to speakers with unknown accents or peculiar pronunciation.

Expanding the system to include more STM levels of representation and, consequently, more LTM connection matrices between them, allows for the construction of a multilevel system with higher-order levels chunking, or grouping, lower-order patterns and their temporal order (Figure 7). The LTM connections between successive levels learn to encode the relative importance and the temporal order of features (i.e., patterns of activation) at one level and represent them as categories at the immediately following level (Grossberg & Stone, 1986). It must be noted that this automatic context-dependent scaling property of ART networks is very important for modeling speech perception where many acoustic cues are often present, but the importance of each depends on its current acoustic–phonetic context. Furthermore, the development of language-specific phonemic representations requires the tuning of sensory features according to the properties that are salient for a particular language (Jusczyk, 1993, 1997).

[4] The well-known phenomenon of phonemic restoration (Samuel & Ressler, 1986; Warren, 1970), whereby listeners perceive speech in which a portion has been substituted by noise as if it were intact, does not fall under this notion of hallucination because the noise replacing the speech signal is compatible with the perceived phoneme, even though not quite specifying it.

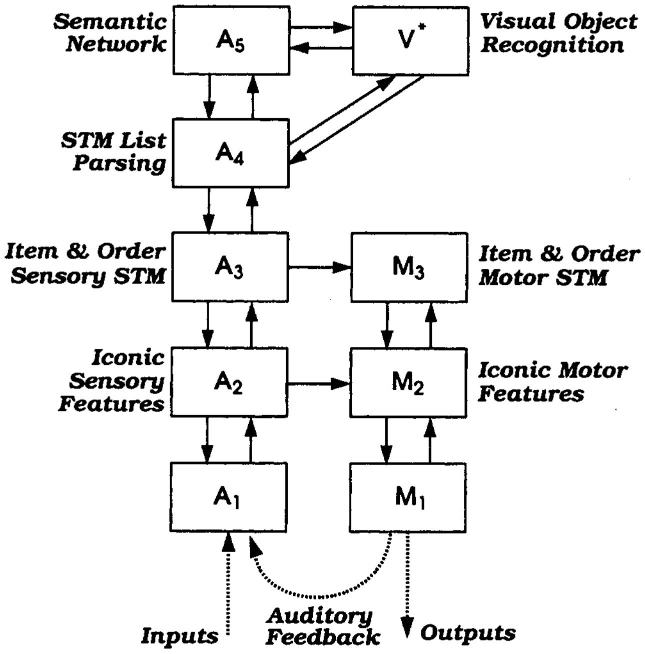

Figure 7. A "macrocircuit governing self-organization of recognition and recall processes" (after Grossberg, 1987, p. 52; Grossberg & Stone, 1986, p. 59). Auditory processes A_i and motor processes M_i self-organize at several levels to encode linguistic features, phonetic categories and phonemes, items (syllables), and lists (words), as required for speech perception and production. Short-term memory (STM) patterns at successive stages affect each other through adaptive filters with self-organized long-term memory traces. Codes stabilize through learning, and associative maps are formed between encoded invariants at the auditory and motor hierarchies. Note that categories, items, and lists need not necessarily correspond to traditional notions of phonemes, syllables, and words. Furthermore, category codes are represented by distributed patterns of activation over entire layers and not locally by single specialized nodes. From "Competitive Learning: From Interactive Activation to Adaptive Resonance," by S. Grossberg, 1987, *Cognitive Science, 11,* p. 52. Copyright 1987 by the Cognitive Science Society. Reprinted with permission.

The context-dependent importance of such features may be exploited by a syllabic grouping at a subsequent level of an ART network.

ART networks of this type have been shown to form stable category codes from input patterns and temporal sequences. Such networks, however, cannot encode embedded and overlapping patterns while maintaining their stability. This realization led to the development of *masking field* networks (Cohen & Grossberg, 1986, 1987; Grossberg, 1986), in which longer patterns are always favored over shorter patterns. If longer patterns were left to compete (without advantage) with old short patterns embedded in them, long words made up of strings of short words could never form stable representations. For example, a word like *catalog* would not be able to overcome the activation of the words *cat, a,* and *log* and establish itself as a unique new category. However, masking fields cannot yet self-organize to form stable category codes like the ART networks, and therefore the long-list advantage must be hardwired into the system.[5]

Articulatory-Phonetic Categories

The development of phonetic categories in infants is a stage necessary for language development that is still not well understood. Gupta and Mozer (1993; Gupta, 1994) proposed a connectionist model that simulates the development of phonological representations through development of attractor states. The model also shows the loss of sensitivity to nonnative contrasts found in 10-month-old infants (Werker & Tees, 1983, 1984; see recent reviews on children's phonetic reorganization in Werker, 1993, 1994). A syllabic nature of phonological representations is a basic assumption of this work that is based on evidence for greater accessibility (in infants) of the syllable than of the phoneme as a perceptual unit (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988; Bertoncini & Mehler, 1981; see recent reviews by Jusczyk, 1997, and Eimas, 1997). A part of the network's success in preference for syllabic units is the result of a gating component in the input to the network. The most interesting demonstration is the gradual loss of phonetic contrasts that are not salient for phonemic categorization when the attractor states are formed.

Additional useful insights on learning and adaptation may be found in models of speech production development, in which the acoustic environment is hypothesized to lead to articulatory motor synergies, which in turn lead to vocal tract configurations that produce sounds gradually approaching those of one's native language. One such model was sketched by Grossberg (1986, pp. 253–257). Recently, an implementation of this idea was developed by Guenther (1994, 1995b) that uses random babbling as a way to explore the articulatory space and its mapping onto auditory space. The model is called Directions (in orosensory space) Into Velocities of Articulators (DIVA) and in its initial formulation used orosensory targets for the sounds. Orosensory targets were chosen over often-proposed muscular or articulatory arrangements to account for the motor equivalence observed in speech production (in that several distinct configurations of individual articulators may lead to the same vocal tract shape, and thus to the same sound). The model was implemented as an adaptive neural network that learned the mapping from phonetic specification to articulatory motor commands through a prewired speech recognition module, training associations between phonetic and orosensory space, and between orosensory and articulatory space (see Figure 8). Subsequent versions of the model (Guenther, 1995a; Johnson & Guenther, 1995) were modified to use acoustic (as opposed to orosensory) targets for the sounds, on the basis of recent evidence that sometimes widely different vocal tract configurations may be used for the production of the same phoneme (Espy-Wilson & Boyce, 1993; Perkell, Matthies, & Svirsky, 1994; Perkell, Matthies, Svirsky, & Jordan, 1994). Preservation of the orosensory mapping allowed the model to capture motor-equivalent compensatory ar-

[5] Nigrin (1990, 1993) developed a self-organized neural network (SONNET) that combined the desirable properties of ART and masking fields. The use of networks like SONNET in the context of speech perception has not yet been investigated. Given the pervasiveness of pattern embedding in words, the combinatorial characteristics of morphology, and the fact that morpheme patterns must be mapped to stable, learned category codes, the architecture of SONNET may be a fruitful approach to modeling word learning and recognition.

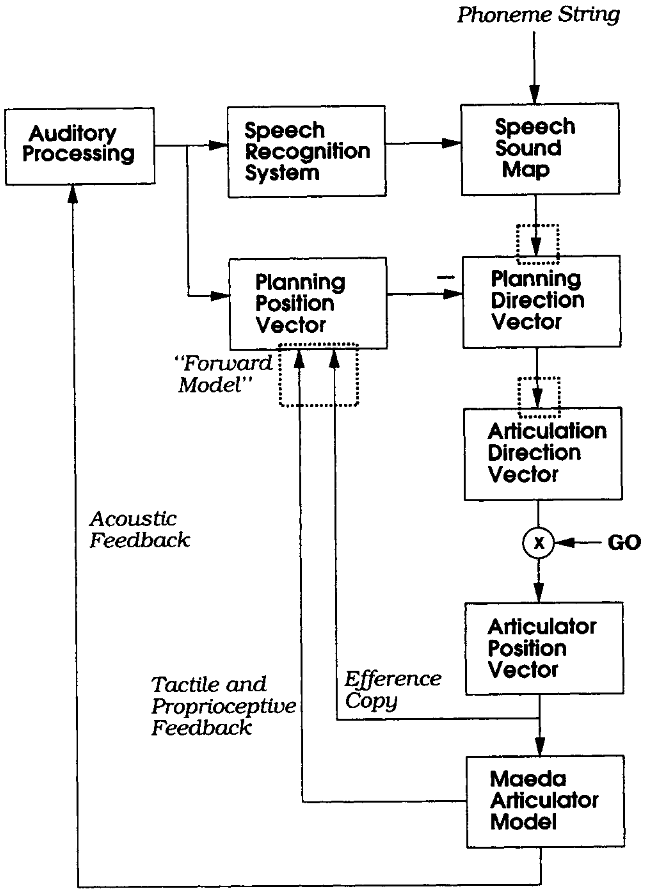

Figure 8. Schematic diagram of the Directions Into Velocities of Articulators (DIVA) model developed by Guenther (1994, 1995b), as revised by Guenther (1995a) and Johnson and Guenther (1995), to model phonetic–articulatory development through random babbling and mapping of auditory space onto articulatory direction space. Learned mappings are indicated by enclosing dotted lines. The most recent version of DIVA is illustrated here, which uses auditory targets as opposed to the orosensory targets of the earlier versions. The orosensory mappings may need to be reincorporated to account for compensatory articulator motion in cases of external perturbation. From "A Modeling Framework for Speech Motor Development and Kinetic Articulator Control," by F. H. Guenther, 1995, in *Proceedings of the XIIIth International Congress of Phonetic Sciences* (Vol. 2, p. 93), Stockholm, Sweden: KTH and Stockholm University. Copyright 1995 by KTH. Reprinted with permission.

ticulation in the case of external perturbation even in the absence of auditory feedback, in agreement with human production data.

Two issues raised by the DIVA model are closely related to the problem of the development of speech perception. First, a major theoretical assumption of DIVA was that the orosensory (initially) and auditory (more recently) targets were not single points in their corresponding space, but convex regions capturing the range along which any one parameter (i.e., feature) is allowed to vary for a given sound. This *convex region hypothesis,* along with the associative learning technique, led to each phoneme's full featural (orosensory or auditory) specification, but with varying importance for each feature. In contrast to underspecified models (discussed below), where only the articulators that are necessary for a phoneme's production are specified, the convex region in target space defines allowable positions for all degrees of freedom. In contrast to fully specified models, where all articulators' positions are defined exactly for every phoneme, DIVA's representation allows greater variability in features whose change would not alter the produced sound's category. This approach led to an impressive fit to experimental data pertaining to anticipatory and carryover coarticulation, speech rate effects on articulator velocity and movement distance, and vowel reduction in fast speech and in languages with sparse vowel spaces (Guenther, 1995b). The implications for speech perception models may be to abandon template matching and the notion of idealized spectra for phonetic identification and to explore instead the acoustic space that can be produced by a vocal tract, assigning entire regions to phonetic categories.

The second important contribution of DIVA that is relevant to this article concerns the nature of the representations in the orosensory and articulatory levels. In particular, instead of coding vocal tract configuration and articulatory positions, DIVA codes direction of change in vocal tract configuration that maps onto articulator movement (cf. the notion that the units of speech are dynamic articulatory actions and not neutral, static constructs; cf. Browman & Goldstein, 1995, p. 181). For example, reducing the lip aperture would map onto upward jaw movement, downward movement of the upper lip, and upward movement of the lower lip. Thus, motor equivalence is modeled very accurately even in situations never encountered during learning, such as simulated bite-block and lip-perturbation experiments (Guenther, 1994). The relevance of this hypothesis for speech recognition is in the importance of concentrating on the dynamic (i.e., time-varying) properties of acoustic representations, looking for targets that are being approached, as opposed to concentrating on static regions of minimal spectral change. Further support for the importance of time-varying spectral portions comes from studies showing higher discriminability of such portions (Furui, 1986; Lindblom & Studdert-Kennedy, 1967) and from the well-known undershoot effect (Lindblom, 1963) frequently observed in natural speech, whereby acoustic targets are approached but not entirely reached. With respect to the development of speech perception, in particular, there is evidence that children are much more dependent on spectral changes relative to static portions than are adults (Nittrouer, 1992; Nittrouer, Crowther, & Miller, 1998; Wallet & Carrell, 1983).

Conclusion

In addition to learning the auditory–articulatory map and the structure of phonemic categories, a complete account of speech perception development must incorporate learning at several levels. For example, learning new words must be designed to be an integral part of the model, but so far only ART includes an account of how such learning might be accomplished (but not how long words could be learned or favored over their shorter component words). The ease of word learning in humans and the flexibility of the lexicon, both in production and in perception, strongly indicate that learning words is, in a way, part of perceiving them. Consider also the need for a complete model to account for nonwords. Nonwords must be completely analyzed and decomposed into their constituent segments (in terms of which the lexicon is described) before it is possible to decide that they are not valid words. Then, their existing description should be immediately usable if they are judged to be unknown words rather than illegal segment sequences. The way to incorporate the novel representations into the lexicon should be similar to the general one-shot learning mechanism that has been observed in many domains of human behavior. Unfortunately, in most existing connectionist models, learning new things usually involves multiple presentations of them, along with repetitions of the already stored data. This is clearly an unacceptable formulation for a novel-word learning model.

Phonological Representation and Computation

It has been noted that, in the course of speech perception, item/type representations of stored knowledge (e.g., phonological structure and lexical form representations) must come in at some point to replace pattern/token representations of sensory processed input (i.e., activation patterns in phonetic feature space). New words, for example, are likely learned by adults on the basis of an already known phonemic inventory, not only by principles of storage economy (which may not be applicable) but also for a parsimonious account of morphological and phonological processes. The nature of the phonological representation(s) and the processes that transform these representations to and from others, such as acoustic/phonetic patterns, lexical items, and motor plans, remain the subject of much speculation. Fortunately, these issues have recently attracted some interest and commendable modeling efforts are under way. As discussed below, some conceptual leaps may be necessary before more exciting results can be obtained, at least with respect to the representation of phonetic features and the dynamical (vs. computational) treatment of phonological processes.

Lexical Underspecification

One view of lexical representation (as well as other aspects of grammar that can be described by systems of distinctive features) holds that properties that are predictable or not distinctive are simply left unspecified (see Archangeli, 1988, for an overview). For example, in English, a vowel is not specified for nasality because there are no English words differing solely on the basis of vowel nasality, and, moreover, this feature is typically assimilated from the following consonant, such that vowels preceding a nasal consonant (e.g., /n/) are realized as [+nasal], whereas vowels preceding nonnasal consonants are not nasalized. More importantly, there are default values for features that are considered predictable, and thus not lexically specified, whereas nondefault values for the same features are specified. An important implication of this postulated representational scheme for word recognition is that a matching asymmetry would be expected: A feature that is not specified lexically would match any input value, whereas a feature that is specified would match (and mismatch) accordingly.

Lahiri and Marslen-Wilson (1991) and, more recently, Nix, Gaskell, and Marslen-Wilson (1993), provided evidence for an underspecified internal representation of lexical items by demonstrating such a matching asymmetry between phonetic features. In Bengali, an Indic language in which nasality is distinctive for vowels, a nasal surface vowel (i.e., in the speech input) was interpreted in a gating task by the listeners as underlyingly either nasal or oral. In contrast, an oral surface vowel was never interpreted as underlyingly nasal. This asymmetry was explained by Lahiri and Marslen-Wilson (1991) on the basis of an underspecified lexical representation in which the feature [+nasal] must be included in the lexicon, but the opposite [−nasal], being the default, is redundant and thus not specified. Therefore, a [+nasal] in the speech input will match a [+nasal] in the lexicon, and will not create a mismatch where no nasality is specified. Likewise, a [−nasal] in the input will mismatch a [+nasal] in the lexicon and will neither match nor mismatch an unspecified representation.

In a similar vein, Nix et al. (1993; see also Gaskell & Marslen-Wilson, 1994, 1996) found that English listeners were likely to interpret a surface labial or velar consonant as coronal in the context of a following labial or velar segment, respectively, but a surface coronal would always be perceived as an underlying coronal. For example, the utterance [leɪk] could be perceived as either *lake* or *late* before [kruz] (*cruise*), but the utterance [leɪt] could only be perceived as *late* regardless of context. (Phonological rules of place assimilation describe these phenomena in more detail.) On the basis of the more recent findings, however, Marslen-Wilson, Nix, and Gaskell (1995) argued that the underspecified representation alone cannot account for the data and that a process of phonological inference is necessary to determine the viability of the transformations given the context (which might span a word boundary).

The claim for an underspecified representation is very attractive and theoretically grounded, but it may be hard to distinguish from alternative theories postulating multiple, fully specified representations. Stevens (1993, 1995), for example, proposed marking features in the (fully specified) lexicon as modifiable, and Klatt (1980) proposed storing several alternative spectral templates for each lexical item to cover the range of variation. Functional underspecification may be the result of any of these implementations. Alternatively, the convex regions hypothesis shown to work for articulatory compensation (Guenther, 1994, 1995b) might work equally well in the perceptual process, especially given an auditory–articulatory mapping as described earlier. In particular, each item (which may be a word or a syllable) can be represented in all dimensions not with point values, but with ranges of allowable variation. The ranges may be context-dependent so that the temporal patterns formed through self-organization differ according to prior and following patterns. These differences may be specific to one dimension or may involve many. For example, place information for syllable-final stop consonants may be allowed to vary between velar and labial. The variation would be in the direction that minimizes the distance between the current and the following region, which specifies the place of the following consonant (cf. Guenther, 1995b, on modeling coarticulation). The reason a region may be more attractive than an unspecified free range is that radical underspecification does not put any restriction on the range of variation of an unspecified feature, and that may not be always desirable. For example, in Korean and Japanese there is no distinction between what English listeners hear as /r/ and /l/, so the corresponding l/r phoneme might be unspecified for the features that distinguish the two in English. In a region representation, however, there would be a context-dependent allowable range of the relevant acoustic–phonetic specification (including, e.g., the third formant frequency), naturally accounting for any syllable-based variation as well. Regions may not offer a functional advantage over multiple specifications but they constitute a more coherent and unified approach to lexical and sublexical representations and may be more naturally implemented in connectionist models.

Phonological Representation and Inference

Gaskell, Hare, and Marslen-Wilson (1995) trained a recurrent neural network to recognize phonetic specifications of labial and velar consonants as underlyingly coronal in the appropriate phonological contexts (e.g., [rɛgkɒr] as *red car* and [gʊbbɔɪ] as *good boy*). They used a network similar to that of Shillcock et al. (1991, see Figure 4, right), again trained to identify the preceding, current, and next phoneme (in phonetic feature specification), given the phonetic feature specification of the current input. According to the human data, the network should learn to interpret surface labials and velars as possible underlying coronals when the following segments were labial or velar, respectively. In addition, if the network is to develop an underspecified phonological representation, it must also learn not to do two things. First, the network must never interpret surface coronals as noncoronals, regardless of context. Second, the network must never interpret a surface velar as an underlying coronal in the context of a following labial and, conversely, never interpret a surface labial as an underlying coronal in the context of a following velar. For example, the network should never interpret [leɪkpraɪd] as *late pride.*

The results of the simulations showed that the network basically learned the autoassociation task between input and output features and that labial and velar input consonants mapped onto coronal output consonants in some cases. Unfortunately, because of the nature of the featural specification, the network failed to pick up exactly how critical the single-feature difference between labial and velar was, so it sometimes responded with an underlying coronal segment in cases of surface labial or velar segments in the opposite context, that is, a following velar or labial respectively. However, the mapping from noncoronal to coronal in different-place contexts was not as frequent as in the same-place contexts. The network sometimes (11% of the time) mapped noncoronals to coronals in other following contexts (e.g., a vowel). Further experimentation with high-frequency words showed that the network represented longer sequences in its recurrent hidden layer (cf. the dynamic-net model of Norris, 1992) and took word identity into account when deciding whether to interpret a surface noncoronal segment as coronal or not.

The two main drawbacks of featural representations such as the one used by Gaskell et al. (1995) are (a) the lack of structure (or affinity relations) and (b) the inability to express continuity (as underlying articulatory actions). Structure is necessary to account for the common featural groupings of phonological transformations and the ensuing generalizations (cf. the notion of underspecification considered in the context of feature geometries and autosegmental phonology; Kenstowicz, 1994, pp. 506–513). Continuity is indispensable in making apparent the nature of assimilations, which can be thought of more naturally as anticipation or perseveration of a gesture (cf. Browman & Goldstein, 1995). In addition to these representational deficiencies, the dissociation between the signified in the input and output layers may be the source of much difficulty in modeling transformational processes, as explained below.

Structure in Phonetic Feature Space

Articulatory (Browman & Goldstein, 1989) and independent theoretical phonological considerations (Clements, 1985) have converged on the development of a structured featural description in which phonetic features, being a direct consequence of articulatory movement, exhibit certain interdependencies that reflect the articulatory anatomy and motor planning (Fowler, 1995). In other words, phonological changes are clustered so that linguistic features that are primarily controlled by physically coupled articulators tend to be affected in unison. In the course of learning to speak and understand a language, ensuing associations develop between physical constraints in articulation and frequent acoustic patterns.

One may think of the mapping between auditory space and articulatory motor control as defining regions of tolerance for each phonetic segment that result from a language-specific optimization of articulatory motor planning. In a self-organized map that develops through exposure to a speech environment, such dependencies must automatically evolve as part of the acoustic–motor structure. In order for this organization to occur, the system must be left to discover the denser regions in articulatory and in auditory space, as well as the mapping between the two, given a primitive feature sensory input (i.e., not of linguistic features but of acoustic properties). Demonstrations that such a developmental process is possible are given by models such as DIVA, the predictions of which closely follow findings on children's speech development.

In a featural representation, it is essential that information about what goes with what be preserved, so that phonological effects are naturally accounted for. In a model that learns its representations from raw auditory input through exposure to endless speech samples, and perhaps with an articulatory loop component, it may be unnecessary to arrange the input features in any structured way. In a case, however, where phonological phenomena are to be accounted for, but the phonetic input is presented preprocessed ad hoc, one must make featural geometry explicit by design. For example, features related to tongue positioning for consonant articulation are likely to function in unison, whereas those that specify vowel roundness and duration should not be allowed to participate in the same phenomena.

Continuity and Identity

Assimilation processes, like the one studied by Marslen-Wilson et al. (1995), are usually viewed as features spreading in time even though they are often formalized as symbol-rewrite rules. Recent advances in articulatory phonology place great importance on the continuity of articulator motion. Assimilatory processes, such as the one leading to the place change under study, are described naturally when the relationships between nodes in a structured feature hierarchy and the continuous identity of each node-feature are preserved. In the context of place assimilation modeling, continuity is of the essence because it makes a substantial (if only conceptual) difference in what the learned associations are. Associating specific input patterns to arbitrary output patterns is likely to sidestep the most important issue in feature assimilation modeling, namely, that the presumed continuity in underlying (articulatory) action gives rise to the observed (phonetic) pattern and as such it may have to be "undone" (phonologically decoded).

In addition to within-layer identity, or continuity, preservation of between-layer identity has been a subject of some discussion in the connectionist literature. The idea is that the network must somehow represent that the features at the input layer are essentially the same thing as the features at the output layer. A concern about neural networks that are trained to map input to output vectors stems from the fact that there is nothing in a network's representation to indicate that the nodes representing, for instance, the output place feature have anything in common with the nodes representing input place features. For a network like the one of Marslen-Wilson et al. (1995), there was just greater variability in the association between some nodes at one layer and other nodes at another layer.

The problem of using structurally unrelated nodes to represent the same kind of information at two different stages in a transformation process is closely related to the criticism on the lack of "preservation of stem and affix" (Pinker & Prince, 1988, p. 108) in an interactive activation model of the past tense (Rumelhart & McClelland, 1986). Specifically, context-dependent phoneme nodes were used in that model to represent a present-tense verb at the input layer and its past-tense form at the output layer. As far as the network was concerned, there was only a mapping from one set of arbitrary pattern activations to another, without any representation of the fact that, in most cases, a part of the input pattern was the same thing as a part of the output pattern, that is, the verb stem. In essence, the regular pattern of transformation includes an autoassociative component and a context-dependent suffixation. Symbolic treatments of morphological and other transformations involve variables, which may stand for certain kinds and inherit all properties of the things they stand for in all instantiations of their representation. For example, denoting a verb stem with s and the standard /d/ past-tense suffix with p, we may write a general rule of the form $[past : s \rightarrow s + p]$, where the same s is present on both sides of the transformation rule.

Fodor and Pylyshyn (1988) made similar remarks regarding the ability of neural networks to exhibit this rule-like systematicity of cognitive and linguistic processes. In a symbolic context, learning once that $s \rightarrow s + p$ is enough for all existing tokens with the "stem of regular verb" property to take the place of s in both sides of the equation. Moreover, p remains the same for all cases of s. In connectionist implementations so far, this systematicity evidently does not arise from the learned mapping. On the other hand, the inability of traditional computational accounts to adequately describe linguistic behavior outside of few well-prescribed (and perfectly systematic) domains suggests that an entirely different approach may be most beneficial. Specifically, recent progress in dynamical systems modeling has shed new light on our understanding of symbols and computations. Recurrent neural networks, being a powerful kind of dynamical system that can be trained to implement any mapping, may arguably constitute our current most promising option. Nearly all models discussed in this article are in fact recurrent networks, and each has particular strong and weak points. None has so far succeeded to stand up to the criticisms of Fodor and Pylyshyn (1988), but the alternatives are far from having been fully explored. More attention to the structure of representations and the learning processes is necessary before connectionist modeling can reach its mature stage in the area of speech perception.

Dynamics Underlying Symbolic Systems

In a dynamical model of psychological processes, there need not be an explicit symbolic representation on which computational processes apply. Kolen (1994) made a case of demonstrating that ascribing computation to a dynamical system is, to a large extent, subjective. That is, unless the performance measures are translated into competence symbols, there can be no computation. Given a time-varying system of local interactions (brain) that produces overt measurables (behavior) and possesses a certain degree of plasticity to strengthen connections between associated events, interactions with the environment lead to a self-organized internal structure that closely mirrors that of the environment. The system's states that are clustered enough to be discernible can be translated into discrete entities (symbols), and the processes that take the system from one state to the other can be then formalized as computations. In the domain of speech production and perception, clusters (ranges) of numerical values of articulatory parameters (gestures) are likewise mapped onto what Browman and Goldstein (1995) called the "macroscopic structure of contrastive categories" (p. 184), that is, the features and phonemes of phonological theory.[6]

Dynamical systems, such as recurrent networks, that exhibit symbolic behavior are neither symbolic systems nor implementations thereof. Recurrent networks operating in real time are the models, even if superimposing a symbolic structure on top of them helps us to see the generalizations better.[7] Recurrent neural networks are dynamical systems capable of behavior of arbitrary complexity, including the implementation of symbolic systems. Pinker and Prince (1988) and Fodor and Pylyshyn (1988) gave connectionism only implementational status, that is, that connectionism models the neural substrate underlying computational systems best described by symbols and rules. However, connectionist models can potentially account not only for all symbolic behavior but also for all departures from it, which symbolic modelers rush to characterize as performance limitations. Surely, much work needs to be done before such claims can be taken literally. It remains to be demonstrated that a functioning system can exhibit human-like systematicity consistently without explicit symbolic computations.

[6] Articulatory parameters are likely unnecessary and perhaps irrelevant in the case of vowel representations because of the variability in vocal tract configurations that can lead to the production of vowels other than /a/, /i/, and /u/. Grouping of acoustic features, such as formant frequencies, can give rise to the discrete linguistic categories while preserving their internal structure (Guenther & Gjaja, 1996). Recent investigations on vowel discrimination have led Aaltonen et al. (1997) to an essentially identical interpretation: "... on the auditory processing level [the listener's experience with spoken language] is organized into clusters of similar, frequently heard speech sounds. The ability to discriminate sounds within the same cluster is impaired ... and at the subsequent phonetic processing level the category limits are adjusted to fit the pattern of clusters" (p. 1102).

[7] Note that Grossberg (1987) also rejects the distinction between a model's low-level mechanism and architecture, and its functional properties.

Importantly, these ideas must be tested at the lexical level. Recent advances in recurrent networks and in self-organizing learning models have greatly expanded our understanding of phonetic perception, organization, and lexical activation, and have only begun to explore the potential of dynamical models in psychological modeling. Beyond lexical activation, however, it is unclear how the dynamical systems view would hold. Norris (1994) had to resort to a copy of TRACE's word level, with temporally aligned localist word nodes, to account for the pervasive effects of context and competition ordinarily encountered in the speech perception literature. ART networks and self-organizing masking fields may constitute an attractive alternative, but, for the time being, no large-scale implementation of multiple levels is available (but see Grossberg et al., 1997, for computer simulations modeling rate-independent speech categorization). Certainly these directions need to be explored in future connectionist models.

Concluding Remarks

Speech perception is a vast, challenging field, many aspects of which have been extensively investigated, though several remain poorly understood. Connectionism offers parallel distributed approaches to modeling, enlightened by both behavioral findings and neurobiological considerations. The recent advances in understanding and using recurrent neural networks have led to impressive progress in speech perception modeling at several levels, including phonetic category formation and identification, phoneme and word activation, and intralexical competition. Although no unified approach has yet emerged that might be useful in building an artificial speech recognition machine, several networks have been used to successfully model human data by matching not only significant differences between experimental conditions (e.g., Gaskell & Marslen-Wilson, 1997; Norris et al., 1995) but sometimes also the time-course of lexical activation (e.g., Allopenna et al., 1998). Still, notably missing from connectionist implementation are prosodic features of speech that have long been implicated in lexical processing and segmentation. Issues of phonological development and word learning are also underdeveloped compared, for example, with lexical activation or the directionality of processing.

The neural and psychological plausibility of the existing models spans a wide range, with the most extensively and successfully applied model, TRACE, being highly implausible, whereas the most neurally rooted theory, adaptive resonance, remains mostly unimplemented in any realistic scale. Researchers have successfully used simple networks with recurrent connections only in the hidden layer to model context-dependent effects without cumbersome temporal representations, but the models suffer from an implausible and inflexible learning scheme. Models of phonetic and articulatory development have pointed at the high potential of self-organized networks in forming appropriate representations and modeling categorical speech effects. Future models must combine the insights learned from different approaches into a coherent multilevel model of speech perception from sound to word.

References

Aaltonen, O., Eerola, O., Hellström, Å., Uusipaikka, E., & Land, A. H. (1997). Perceptual magnet effect in the light of behavioral and psychophysiological data. *The Journal of the Acoustical Society of America, 101,* 1090–1105.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38,* 419–439.

Altmann, G. T. M. (1990). Cognitive models of speech processing: An introduction. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 1–23). Cambridge, MA: MIT Press.

Altmann, G. T. M., & Shillcock, R. (Eds.). (1993). *Cognitive models of speech processing: The second Sperlonga meeting.* Hillsdale, NJ: Erlbaum.

Anderson, J. A., Pellionisz, A., & Rosenfeld, E. (Eds.). (1990). *Neurocomputing 2: Directions for research.* Cambridge, MA: MIT Press.

Anderson, J. A., & Rosenfeld, E. (Eds.). (1988). *Neurocomputing: Foundations of research.* Cambridge, MA: MIT Press.

Anderson, J. A., Rossen, M. L., Viscuso, S. R., & Sereno, M. E. (1990). Experiments with representation in neural networks: Object motion, speech, and arithmetic. In H. Haken & M. Standler (Eds.), *Synergetics of cognition* (pp. 54–69). Berlin: Springer.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84,* 413–451.

Archangeli, D. (1988). Aspects of underspecification theory. *Phonology, 5,* 183–207.

Aslin, R. N., Jusczyk, P. W., & Pisoni, D. B. (1998). Speech and auditory processing during infancy: Constraints on and precursors to language. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology, Vol. 2: Cognition, perception, & language* (pp. 147–198). New York: Wiley.

Bard, E. G., & Shillcock, R. C. (1993). Competitor effects during lexical access: Chasing Zipf's tail. In G. T. M. Altmann & R. Shillcock (Eds.), *Cognitive models of speech processing: The second Sperlonga meeting* (pp. 235–275). Hillsdale, NJ: Erlbaum.

Bard, E. G., Shillcock, R. C., & Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics, 44,* 395–408.

Bertoncini, J., Bijeljac-Babic, R., Blumstein, S. E., & Mehler, J. (1987). Discrimination in neonates of very short CV's. *The Journal of the Acoustical Society of America, 82,* 31–37.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representation of speech sounds. *Journal of Experimental Psychology: General, 117,* 21–33.

Bertoncini, J., & Mehler, J. (1981). Syllables as units of infant speech perception. *Infant Behavior and Development, 4,* 247–260.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience, 2,* 32–48.

Blomberg, M., Carlson, R., Elenius, K., & Granström, B. (1986). Auditory models as front ends in speech-recognition systems. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 108–122). Hillsdale, NJ: Erlbaum.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6,* 201–251.

Browman, C. P., & Goldstein, L. (1995). Dynamics and articulatory phonology. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 175–193). Cambridge, MA: MIT Press.

Carpenter, G. A., & Govindarajan, K. K. (1993). *Speaker normalization methods for vowel recognition: Comparative analysis using neural network and nearest neighbor classifiers* (Tech. Rep. No. CAS/CNS-93-039). Boston, MA: Boston University Center for Adaptive Systems and Department of Cognitive and Neural Systems.

Carpenter, G. A., & Grossberg, S. (Eds.). (1991). *Pattern recognition by self-organizing neural networks.* Cambridge, MA: MIT Press.

Carpenter, G. A., & Grossberg, S. (1995). *Adaptive resonance theory: Self-organizing networks for stable learning, recognition, and prediction* (Tech. Rep. No. CAS/CNS-95-017). Boston, MA: Boston University Center for Adaptive Systems and Department of Cognitive and Neural Systems.

Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain.* Cambridge, MA: MIT Press.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook, 2,* 225–252.

Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 551–563.

Cohen, M. A., & Grossberg, S. (1986). Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short-term memory. *Human Neurobiology, 5*(1), 1–22.

Cohen, M. A., & Grossberg, S. (1987). Masking fields: A massively parallel neural architecture for learning, recognizing, and predicting multiple groupings of patterned data. *Applied Optics, 26*(10), 1866–1891.

Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language, 32,* 193–210.

Cutler, A., Mehler, J., Norris, D., & Seguí, J. (1987). Phoneme identification and the lexicon. *Cognitive Psychology, 19,* 141–177.

Dupoux, E. (1993). The time course of prelexical processing: The syllabic hypothesis revisited. In G. T. M. Altmann & R. Shillcock (Eds.), *Cognitive models of speech processing: The second Sperlonga meeting* (pp. 81–114). Hillsdale, NJ: Erlbaum.

Eimas, P. D. (1997). Infant speech perception: Processing characteristics, representational units, and the learning of words. In R. L. Goldstone, D. L. Medin, & P. G. Schyns (Eds.), *The psychology of learning and motivation: Perceptual learning* (Vol. 36, pp. 127–169). New York: Academic Press.

Eimas, P. D., Marcovitz Hornstein, S. B., & Payton, P. (1990). Attention and the role of dual codes in phoneme monitoring. *Journal of Memory and Language, 29,* 160–180.

Eimas, P. D., & Nygaard, L. C. (1992). Contextual coherence and attention in phoneme monitoring. *Journal of Memory and Language, 31,* 375–395.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171,* 303–306.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179–211.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Elman, J., & McClelland, J. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 360–385). Hillsdale, NJ: Erlbaum.

Elman, J. L., & Zipser, D. (1988). Learning the hidden structure of speech. *The Journal of the Acoustical Society of America, 83*(4), 1615–1626.

Espy-Wilson, C., & Boyce, S. (1993). Acoustic differences between "bunched" and "retroflex" variants of American English /r/. *The Journal of the Acoustical Society of America, 95*(5, Part 2), 2823.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28,* 3–71.

Fowler, C. A. (1995). Speech production. In J. L. Miller & P. D. Eimas (Eds.), *Speech, language, and communication* (pp. 29–61). San Diego, CA: Academic Press.

Frauenfelder, U. H., & Peeters, G. (1990). Lexical segmentation in TRACE: An exercise in simulation. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 50–86). Cambridge, MA: MIT Press.

Frauenfelder, U. H., Seguí, J., & Dijkstra, T. (1990). Lexical effects in phonemic processing: Facilitatory or inhibitory? *Journal of Experimental Psychology: Human Perception and Performance, 16,* 77–91.

Furui, S. (1986). On the role of spectral transition for speech perception. *The Journal of the Acoustical Society of America, 80*(4), 1016–1025.

Gaskell, M. G. (1994). *Spoken word recognition: A combined computational and experimental approach.* Unpublished doctoral dissertation, Birkbeck College, University of London.

Gaskell, M. G., Hare, M., & Marslen-Wilson, W. D. (1995). A connectionist model of phonological representation in speech perception. *Cognitive Science, 19,* 407–439.

Gaskell, G., & Marslen-Wilson, W. (1994). Inference processes in speech perception. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society* (pp. 341–345). Hillsdale, NJ: Erlbaum.

Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 144–158.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12,* 613–656.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28,* 501–518.

Gow, D. W., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 344–359.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception & Psychophysics, 38*(4), 299–310.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition in humans and machines, Vol. 1: Speech perception* (pp. 187–294). Orlando, FL: Academic Press.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11,* 23–63.

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 481–503.

Grossberg, S., & Stone, G. (1986). Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance. *Psychological Review, 93,* 46–74.

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics, 72,* 43–53.

Guenther, F. H. (1995a). A modeling framework for speech motor development and kinematic articulator control. In *Proceedings of the XIIIth International Congress of Phonetic Sciences* (Vol. 2, pp. 92–99). Stockholm, Sweden: KTH & Stockholm University.

Guenther, F. H. (1995b). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review, 102,* 594–621.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America, 100,* 1111–1121.

Gupta, P. (1994). Investigating phonological representations: A modeling agenda. In M. C. Mozer, P. Smolensky, D. S. Touretzky, J. L. Elman, & A. S. Weigend (Eds.), *Proceedings of the 1993 Connectionist Models Summer School.* Hillsdale, NJ: Erlbaum.

Gupta, P., & Mozer, M. C. (1993). Exploring the nature and development of phonological representations. In *Program of the Annual Conference of the Cognitive Science Society, Vol. 15.* Hillsdale, NJ: Erlbaum.

Haffner, P., & Waibel, A. (1992). Multi-state time delay neural networks for continuous speech recognition. In *Advances in neural information processing systems* (Vol. 4, pp. 135–142). San Francisco: Morgan Kaufman.

Hertz, J., Krogh, A., & Palmer, R. (1991). *Introduction to the theory of neural computation.* Redwood City, CA: Addison-Wesley.

Intrator, N., & Cooper, L. N. (1992). Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks, 5,* 3–17.

Jakobson, R., Fant, C. G. M., & Halle, M. (1972). *Preliminaries to speech analysis: The distinctive features and their correlates.* Cambridge, MA: MIT Press.

Johnson, D., & Guenther, F. H. (1995). Acoustic space movement planning in a neural network model of motor equivalent vowel production. In *Proceedings of the World Congress on Neural Networks* (pp. 481–484). Washington, DC: International Neural Network Society Press.

Johnson, K., & Mullennix, J. W. (Eds.). (1997). *Talker variability in speech processing.* San Diego, CA: Academic Press.

Jordan, M. (1986). *Serial order: A parallel distributed processing approach* (Institute for Cognitive Science Report No. 8604). University of California, San Diego.

Jusczyk, P. (1993). How word recognition may evolve from infant speech perception capacities. In G. T. M. Altmann & R. Shillcock (Eds.), *Cognitive models of speech processing: The second Sperlonga meeting* (pp. 27–55). Hillsdale, NJ: Erlbaum.

Jusczyk, P. W. (1997). *The discovery of spoken language.* Cambridge, MA: MIT Press.

Kenstowicz, M. (1994). *Phonology in generative grammar.* Cambridge, MA: Blackwell.

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

Kolen, J. F. (1994). *Exploring the computational capabilities of recurrent neural networks.* Unpublished doctoral dissertation, Ohio State University, Columbus.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50,* 93–107.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255,* 606–608.

Lahiri, A., & Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition, 38,* 245–294.

Levy, J., Shillcock, R., & Chater, N. (1991). Connectionist modelling of phonotactic constraints in word recognition. In *International joint conference on neural networks* (pp. 101–106). Singapore: IEEE.

Liberman, A. L., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America, 35,* 1773–1781.

Lindblom, B., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *The Journal of the Acoustical Society of America, 42,* 830–843.

Lippman, R. P. (1989). Review of neural networks for speech recognition. *Neural Computation, 1,* 1–38.

Lively, S. E., & Pisoni, D. B. (1997). On prototypes and phonetic categories: A critical assessment of the perceptual magnet effect in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 1665–1679.

Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics, 39*(3), 155–158.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 122–147). Cambridge, MA: MIT Press.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition, 25,* 71–102.

Marslen-Wilson, W. (Ed.). (1989). *Lexical representation and process.* Cambridge, MA: MIT Press.

Marslen-Wilson, W., Brown, C. M., & Tyler, L. K. (1988). Lexical representations in spoken language comprehension. *Language and Cognitive Processes, 3*(1), 1–16.

Marslen-Wilson, W., Nix, A., & Gaskell, G. (1995). Phonological variation in lexical access: Abstractness, inference, and English place assimilation. *Language and Cognitive Processes, 10*(3/4), 285–308.

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

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

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 576–585.

Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language, 27,* 213–234.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology, 21,* 398–421.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology, 23,* 1–44.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 433–443.

McQueen, J. M., Norris, D., & Cutler, A. (1994). Competition in spoken word recognition: Spotting words in other words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 621–638.

Miller, J. L., & Eimas, P. D. (1995a). Speech perception: From signal to word. *Annual Reviews in Psychology, 46,* 467–492.

Miller, J. L., & Eimas, P. D. (Eds.). (1995b). *Speech, language, and communication.* San Diego, CA: Academic Press.

Mozer, M. C. (1993). Neural network architectures for temporal pattern processing. In A. S. Weigend & N. A. Gershenfeld (Eds.), *Time series prediction: Forecasting the future and understanding the past* (pp. 243–264). Redwood City, CA: Addison-Wesley.

Mullenix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America, 85*(1), 365–378.

Nigrin, A. (1990). *The stable learning of temporal patterns with an adaptive resonance circuit.* Unpublished doctoral dissertation, Duke University, Durham, NC.

Nigrin, A. (1993). *Neural networks for pattern recognition.* Cambridge, MA: MIT Press.

Nittrouer, S. (1992). Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. *Journal of Phonetics, 20,* 351–382.

Nittrouer, S., Crowther, C. S., & Miller, M. E. (1998). The relative weighting of acoustic properties in the perception of [s]+stop clusters by children and adults. *Perception & Psychophysics, 60,* 51–64.

Nix, A., Gaskell, G., & Marslen-Wilson, W. (1993). Phonological variation and mismatch in lexical access. In *Proceedings of the 3rd Eurospeech'93.* Berlin, Germany: European Speech Communication Association.

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

Norris, D. (1990). A dynamic-net model of human speech recognition. In

G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 87–104). Cambridge, MA: MIT Press.

Norris, D. (1992). Connectionism: A new breed of bottom-up model? In R. G. Reily & N. E. Sharkey (Eds.), *Connectionist approaches to natural language processing* (pp. 351–371). Hove, UK: Erlbaum.

Norris, D. (1993). Bottom-up connectionist models of "interaction." In G. T. M. Altmann & R. Shillcock (Eds.), *Cognitive models of speech processing: The second Sperlonga meeting* (pp. 211–234). Hillsdale, NJ: Erlbaum.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1209–1228.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5,* 42–46.

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cognition. *Trends in Cognitive Sciences, 2,* 455–462.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation, 1,* 263–269.

Perkell, J. S., Matthies, M. L., & Svirsky, M. A. (1994). Articulatory evidence for acoustic goals for consonants. *The Journal of the Acoustical Society of America, 96*(5, Part 2), 3326.

Perkell, J. S., Matthies, M. L., Svirsky, M. A., & Jordan, M. I. (1994). Trading relations between tongue-body raising and lip rounding in production of the vowel /i/: A pilot "motor equivalence" study. *The Journal of the Acoustical Society of America, 93,* 2948–2961.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed model of language acquisition. *Cognition, 28,* 73–193.

Pitt, M. A., & Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phonetic identification task. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 699–725.

Pitt, M. A., & Samuel, A. G. (1995). Lexical and sublexical feedback in auditory word recognition. *Cognitive Psychology, 29,* 149–188.

Port, R. F., Cummins, F., & McAuley, J. D. (1995). Naive time, temporal patterns, and human audition. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 339–371). Cambridge, MA: MIT Press.

Quinlan, P. (1991). *Connectionism and psychology.* Chicago: University of Chicago Press.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin, 92,* 81–110.

Repp, B. H. (1983). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 10, pp. 243–335). New York: Academic Press.

Rossen, M. L. (1989). *Speech syllable recognition with a neural network.* Unpublished doctoral dissertation, Brown University, Providence, RI.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing, explorations in the microstructure of cognition, Vol. 2: Psychological and biological models* (pp. 216–271). Cambridge, MA: MIT Press.

Samuel, A. G., & Ressler, W. H. (1986). Attention within auditory word perception: Insights from the phonemic restoration illusion. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 70–79.

Seebach, B. S. (1990). *Evidence for the development of phonetic property detectors in a neural net without innate knowledge of linguistic structure.* Unpublished doctoral dissertation, Brown University, Providence, RI.

Seebach, B. S., Intrator, N., Lieberman, P., & Cooper, L. N. (1994). A model of prenatal acquisition of speech parameters. *Proceedings of the National Academy of Sciences, 91,* 7473–7476.

Shillcock, R. (1990). Lexical hypotheses in continuous speech. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 24–49). Cambridge, MA: MIT Press.

Shillcock, R., Levy, J., & Chater, N. (1991). A connectionist model of auditory word perception in continuous speech. In *Program of the annual conference of the Cognitive Science Society, Vol. 13* (pp. 340–345). Hillsdale, NJ: Erlbaum.

Shillcock, R., Lindsey, G., Levy, J., & Chater, N. (1992). A phonologically motivated input representation for the modelling of auditory word perception in continuous speech. In *Program of the annual conference of the Cognitive Science Society, Vol. 14* (pp. 408–413). Hillsdale, NJ: Erlbaum.

Stevens, K. N. (1993). Lexical access from features. In *Speech communication group working papers* (Vol. 8, pp. 119–144). Cambridge, MA: Research Laboratory of Electronics, Massachusetts Institute of Technology.

Stevens, K. N. (1995). Applying phonetic knowledge to lexical access. In *Proceedings of Eurospeech '95* (Vol. 1, pp. 3–11). Madrid, Spain: European Speech Communication Association.

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

Sussman, J. E., & Lauckner-Morano, V. J. (1995). Further test of the "perceptual magnet effect" in the perception of [i]: Identification and change/no-change discrimination. *The Journal of the Acoustical Society of America, 97,* 539–552.

Tabossi, P., Burani, C., & Scott, D. (1995). Word identification in fluent speech. *Journal of Memory and Language, 34,* 440–467.

Tebelskis, J. (1995). *Speech recognition using neural networks.* Unpublished doctoral dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Thorpe, S. J., & Imbert, M. (1989). Biological constraints on connectionist modeling. In R. Pfeifer (Ed.), *Connectionism in perspective.* Amsterdam: North-Holland.

Waibel, A. (1989). Modular construction of time-delay neural networks for speech recognition. *Neural Computation, 1,* 39–46.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 37,* 328–339.

Wallet, A. C., & Carrell, T. D. (1983). Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants. *The Journal of the Acoustical Society of America, 73,* 1011–1022.

Warren, R. (1970). Perceptual restoration of missing speech sounds. *Science, 167,* 392–393.

Watrous, R. L. (1990). Phoneme discrimination using connectionist networks. *The Journal of the Acoustical Society of America, 87*(4), 1753–1772.

Watrous, R. L., Ladendorf, B., & Kuhn, G. (1990). Complete gradient optimization of a recurrent network applied to /b/, /d/, /g/ discrimination. *The Journal of the Acoustical Society of America, 87*(3), 1301–1309.

Werker, J. F. (1993). Developmental changes in cross-language speech perception: Implications for cognitive models of speech processing. In G. T. M. Altmann & R. Shillcock (Eds.), *Cognitive models of speech processing: The second Sperlonga meeting* (pp. 57–78). Hillsdale, NJ: Erlbaum.

Werker, J. F. (1994). Cross-language speech perception: Developmental change does not involve loss. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from spoken sounds to spoken words* (pp. 93–120). Cambridge, MA: MIT Press.

Werker, J. F., & Tees, R. C. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology, 37,* 278–286.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7,* 49–63.

Windheuser, C., & Bimbot, F. (1993). Phonetic features for spelled letter recognition with a time delay neural network. In *Proceedings of Eurospeech '93* (pp. 1489–1492). Berlin, Germany: European Speech Communication Association.

Zwitserlood, P. (1989). The locus of sentential-semantic context in spoken-word processing. *Cognition, 32,* 25–64.

Received July 10, 1997
Revision received August 4, 1998
Accepted December 10, 1998 ■