The Emergence of Phonology from the Interplay of Speech Comprehension and Production: A Distributed Connectionist Approach

David C. Plaut

Christopher T. Kello

Department of Psychology
Carnegie Mellon University

Center for the Neural Basis of Cognition
Carnegie Mellon University and the University of Pittsburgh

February 3, 1998

To appear in B. MacWhinney (Ed.), *The emergence of language*. Mahweh, NJ: Erlbaum.

How do infants learn to understand and produce spoken language? Despite decades of intensive investigation, the answer to this question remains largely a mystery. This is, in part, because, although the use of language seems straightforward to adult native speakers, speech recognition and production present the infant with numerous difficult computational problems (see Lively, Pisoni, & Goldinger, 1994). First of all, processing speech is difficult both because it is extended in time and because it is subject to considerable variability across speakers and contexts. Moreover, even with an accurate representation of the underlying phonetic content of a heard utterance, mapping this representation onto its meaning is difficult because the relationship of spoken words to their meanings is essentially arbitrary. On the output side, the infant must learn to produce comprehensible speech in the absence of any direct feedback from caretakers or the environment as to what articulatory movements are required to produce particular sound patterns. Finally, the processes of learning to understand speech and learning to produce it must be closely related (although certainly not synchronized; Benedict, 1979) to ensure that they eventually settle on mutually consistent solutions.

In the current work, we formulate a general framework for understanding how the infant surmounts these challenges, and we present a computational simulation of the framework that learns to understand and produce spoken words in the absence of explicit articulatory feedback. Our initial focus is on addressing the relevant computational issues; we postpone consideration of how the approach accounts for specific empirical phenomena until the General Discussion.

The framework, depicted in abstract form in Figure 1, is based on connectionist/parallel distributed

*This research was supported by the National Institute of Mental Health (Grant MH47566 and the CMU Psychology Department Training Grant on "Individual Differences in Cognition") and the National Science Foundation (Grant 9720348). We thank Marlene Behrmann, Brian MacWhinney, Jay McClelland, and the CMU PDP Research Group for helpful comments and discussions. Correspondence may be directed either to David Plaut (plaut@cmu.edu) or to Chris Kello (kello@cnbc.cmu.edu), Mellon Institute 115—CNBC, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh PA 15213–2683.

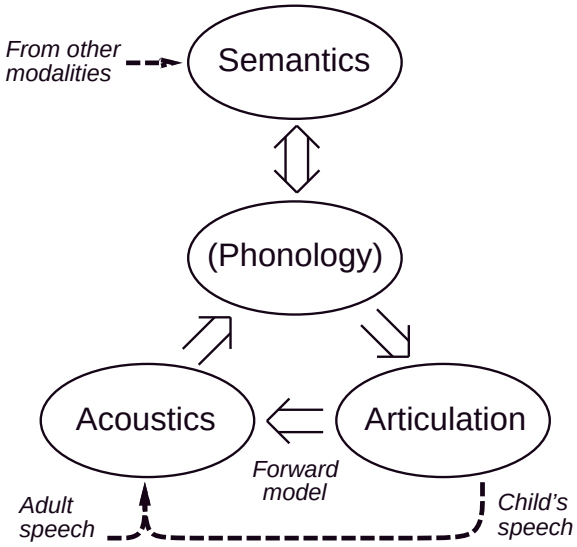

Figure 1. An abstract connectionist framework for phonological development. Ovals represent groups of processing units, and arrows represent mappings among these groups. The intermediate or "hidden" representations which mediate these mappings, and the specific sets of connections among unit groups, are omitted for clarity (see Figure 2 for details). Although "Phonology" is also a learned, hidden representation, and therefore listed in parentheses, it it singled out because it plays a unique role in performing the full range of relevant tasks. The dashed arrows indicate sources of input—either acoustic input from speech generated by an adult model or by the system itself, or semantic input, assumed to be derived from other modalities, such as vision.

processing (PDP) principles, in which different types of information are represented as patterns of activity over separate groups of simple, neuron-like processing units. Within the framework, phonological representations play a central role in mediating among acoustic, articulatory, and semantic representations. Critically, phonological representations are not predefined but are learned by the system under the pressure of understanding and producing speech. In this way, the approach sidesteps the perennial question of what are the specific "units" of phonological representation (see, e.g., Ferguson & Farwell, 1975; Menn, 1978; Moskowitz, 1973; Treiman & Zukowski, 1996). Representations of segments (phonemes) and other structures (onset, rime, syllable) are not built-in; rather, the relevant similarity among phonological representations at multiple levels emerges gradually over the course of development (also see Lindblom, 1992; Lindblom, MacNeilage, & Studdert-Kennedy, 1984; Nittrouer, Studdert-Kennedy, & McGowan, 1989). Also note that the system lacks any explicit structures corresponding to words, such as logogens (Morton, 1969) or "localist" word units (Dell, 1986; McClelland & Rumelhart, 1981). Instead, the lexical status of certain acoustic and articulatory sequences is reflected only in the nature of the *functional* interactions between these inputs and other representations in the system, including semantics (see Plaut, 1997; Van Orden & Goldinger, 1994; Van Orden, Pennington, & Stone, 1990, for discussion). Although the current work focuses on the comprehension and production of single words, the general approach is inherently sequential and, thus, intended to be extensible to higher levels of language processing.

Using distributed representations for words has important—and seemingly problematic—implications for both comprehension and production. In the current formulation, comprehension involves mapping time-varying acoustic input onto a more stable semantic representation (via phonology), whereas production involves generating time-varying articulatory output from a stable phonological "plan" (possibly derived from semantics). The problem is as follows. Distributed connectionist models are strongly biased by *simi-*

larity; because unit activations are determined by a weighted sum of other activations, similar input patterns tend to cause similar output patterns. This property is a boon in most domains, which are largely systematic, because it enables effective generalization to novel inputs (see, e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996). It poses a particular challenge in the current context, however, because of the lack of systematicity between the surface forms of words and their meanings; acoustic/articulatory similarity is unrelated to semantic similarity. This challenge is not insurmountable—distributed connectionist models can learn arbitrary mappings, even for reasonable-sized vocabularies (e.g., a few thousand words; see Plaut, 1997), although they require a large number of training presentations to do so. Learning an unsystematic mapping is even more difficult, however, when the relevant surface information is extended in time. For example, as the initial portion of the acoustic input for a word comes in (e.g., the /m/ in MAN), its implications for the semantics of the word depend strongly on the acoustics for the final portion of the word (e.g., the final /n/; cf. MAT, MAP, MAD, etc.). However, by the time this final portion is encountered, the initial acoustics are long gone.

The obvious (if vague) answer to this problem is "memory." The system must somehow retain earlier portions of the input so that they can be integrated with later portions in order to map a representation of the entire word onto semantics. The critical questions, though, are exactly how this is done and how it is learned. On our view, the answers to these questions are of fundamental importance for understanding the nature of phonological representations.

To solve this problem, we adopt (and adapt) the approach taken by St. John and McClelland (1990, also see McClelland, St. John, & Taraban, 1989) in confronting a similar challenge relating to lexical and syntactic ambiguities in sentence comprehension: the information that resolves a point of ambiguity often comes much later in a sentence. St. John and McClelland trained a simple recurrent network to take sequences of words forming sentences as input, and to derive an internal representation of the event described by the sentence, termed the *Sentence Gestalt*. Critically, the Sentence Gestalt representation was not predefined but was learned based on feedback on its ability to generate appropriate thematic role assignments for the event (via a "query" network). For our purposes, there are two critical aspects of their approach. The first and most straightforward is the use of a recurrent network architecture in which the processing of any given input can be influenced by learned, internal representations of past inputs. The second is more subtle but no less important. From the very beginning of a sentence and for every word within it, the current (incomplete) Sentence Gestalt representation was pressured to derive the correct thematic role assignments of the entire sentence. It was, of course, impossible for the network to be fully accurate at early stages within a sentence, for exactly the reasons we have been discussing: the correct interpretation of the beginning of a sentence depends on later portions which have yet to occur. Even so, the network could be partially accurate—it could at least improve its generation of those aspects of the role assignments that did not depend on the rest of the sentence. In doing so, it was pressured to represent information about the beginning of the sentence within the Sentence Gestalt, thereby indirectly making it available to bias the interpretation of rest of the sentence as necessary. In this way, the "memory" of the system was not any sort of buffer or unstructured storage system but was driven entirely by the functional demands of the task and the interdependence of different types of information.

The same approach can be applied at the level of comprehending single words from time-varying acoustic input. As acoustic information about the very beginning of the word becomes available, and throughout the duration of the word, the system can be trained to activate the full semantic representation of the entire word (which, we assume, is made available from other modalities of input, such as vision). As with sentence-level comprehension, this cannot be done completely accurately, but the network can be pressured to derive whatever semantic implications are possible from the available input to that point (e.g., ruling out some semantic features and partially activating others). Moreover, the network will activate the full representation as soon as the word can be reliable distinguished from all other words (i.e., its uniqueness point; cf. Marslen-Wilson's 1987, Cohort model). This type of processing fits naturally with evidence supporting

the immediacy of on-line comprehension processes (e.g., Eberhard, Spivey-Knowlton, & Tanenhaus, 1995; Sedivy, Tanenhaus, & Spivey-Knowlton, 1995).

Within our framework, phonology (like the Sentence Gestalt) is a learned, internal representation that plays a critical role in mediating between time-varying acoustic input and stable semantic representations (see Figure 1). In particular, we assume that a phonological representation of an entire word builds up gradually over time under the influence of a sequence of acoustic inputs, but that, at every stage, the current approximation is mapped to semantics in parallel. Note that, although the phonological representation encodes the pronunciation of an entire word simultaneously, it must nonetheless retain whatever order and content information in the original acoustic signal is necessary for deriving the correct semantic representation. In this way, learned phonological representations compensate for the detrimental effects of sequential input in learning an unsystematic mapping by integrating and recoding time-varying information into a more stable format. Put simply, phonological representations instantiate the "memory" necessary to map acoustics to semantics.

Analogous issues arise in the production of articulatory sequences from a stable phonological "plan" (Jordan, 1986). In this case, the system must keep track of where it is in the course of executing an articulation so as to apply the appropriate contextual constraints. As an extreme example, consider the word Mississippi /mɪsɪsɪpi/. Without clear information about having completed both the second and third syllables, the system might very well continue on with /mɪsɪsɪsɪsɪ.../. Thus, both comprehension and production require a recurrent network architecture that is capable of integrating information over time in mapping time-varying surface forms to and from more stable internal (phonological) representations.

There is, however, a more fundamental problem to solve regarding production. This problem stems from the fact that the environment provides no direct feedback concerning the appropriate output patterns for production (i.e., the articulations necessary to produce certain sounds). In a sense, the system must discover what sequences of articulatory commands produce comprehensible speech. A critical assumption in the current approach is that the process of learning to generate accurate articulatory output in production is driven by feedback from the comprehension system—that is, from the acoustic, phonological, and semantic consequences of the system's own articulations (also see Markey, 1994; Menn & Stoel-Gammon, 1995; Perkell, Matthies, Svirsky, & Jordan, 1995; Studdert-Kennedy, 1993). Deriving this feedback is made difficult, however, by the fact that, whereas the mapping from articulation to acoustics is well-defined, the reverse mapping is one-to-many (Atal, Chang, Mathews, & Tukey, 1978). That is to say, essentially the same acoustics can be produced by many different articulatory configurations. For example, if no exhalation is allowed, then any static position of the tongue and jaw will result in silence. Silence maps to many possible articulatory states, but each of those articulatory states maps only to silence. From the perspective of control theory, the mapping from proximal domain (articulation) to the distal domain (acoustics) is termed the *forward* mapping, whereas the reverse is termed the *inverse* mapping (see Jordan, 1992, 1996). When the inverse mapping is many-to-one, as in this case, it constitutes a "motor equivalence" problem. This problem must be solved if the link between articulation and acoustics is to support the acquisition of speech production.

Our solution to the motor equivalence problem is based on a computational method developed by Jordan and Rumelhart (1992) (see Markey, 1994, for an alternative approach based on reinforcement learning). The method capitalizes on the fact that, although one cannot deterministically translate distal to proximal *states*, one can translate distal to proximal *errors*.[1] This is accomplished by first learning an internal model of the physical processes that relate specific articulations to the acoustics they produce (recall that the articulation-to-acoustics mapping, although complicated, is well-defined). This *forward model* must be invertible in the sense that acoustic error for a given articulation can be translated back into articulatory error—a natu-

[1] Although we will describe Jordan and Rumelhart's (1992) method in terms of error-correcting learning, it is applicable to any supervised learning framework.

ral instantiation of this would be back-propagation within a connectionist network (Rumelhart, Hinton, & Williams, 1986). Such a model can be learned by executing a variety of articulations, predicting how they each will sound, and then adapting the model based on the discrepancies between these predictions and the actual resulting acoustics. In the infant, we assume that a articulatory-acoustic forward model develops primarily as a result of canonical and variegated babbling in the second half of the first year (Fry, 1966; Oller, 1980; see Vihman, 1996, for review, and Houde, 1997; Wolpert, Ghahramani, & Jordan, 1995, for empirical evidence supporting the existence of forward models in human motor learning). Note that the strong reliance on learning within the current approach contrasts sharply with accounts in which the perceptuomotor associations involved in speech production are assumed to be specified innately (e.g., Liberman & Mattingly, 1985).

The learned forward model plays a critical role within our framework by providing the necessary feedback for learning speech production (also see Perkell et al., 1995). Specifically, the forward model is used to convert acoustic and phonological feedback (i.e, whether an utterance sounded right) into articulatory feedback, which is then used to improve the mapping from phonology to articulation. We assume that learning to produce speech takes place in the context of attempts to imitate adult speech, and attempts to produce intentional utterances driven by semantic representations. In imitation, the system first derives acoustic and phonological representations for an adult utterance during comprehension (see Figure 1). It then uses the resulting phonological representation as input to generate a sequence of articulatory gestures. These gestures, when executed, result in acoustics which are then mapped back onto phonology via the comprehension system. The resulting discrepancies between the original acoustic and phonological representations generated by the adult and those now generated by the system itself constitute the error signals that ultimately drive articulatory learning. In order for this to work, however, these distal errors must be converted to proximal errors (i.e., discrepancies in articulation). This is done by propagating phonological error back to acoustics and then back across the forward model (which mimics the actual physical processes that produced the acoustics from articulation) to derive error signals over articulatory states. These signals are then used to adapt the production system (i.e., the mapping from stable phonological representations onto articulatory sequences) to better approximate the acoustics and phonology generated by the adult. Intentional naming involves similar processing except that the initial input and the resulting comparison are at the level of semantics rather than at acoustics and phonology.

In summary, in the current work we develop a framework, based on connectionist/parallel-distributed processing principles, for understanding the interplay of comprehension and production in phonological development. Comprehension is instantiated as a mapping from time-varying acoustic input onto a more stable internal phonological representation of the entire word which is mapped simultaneously onto its semantic representation. In production, the same phonological representation serves as the input or "plan" for generating time-varying articulatory output. Articulatory feedback is not provided directly but is derived by the system itself from the consequences of self-generated speech using a learned forward model of the physical processes relating articulation to acoustics.

These processes can be instantiated in terms of four types of training experiences: 1) *babbling*, in which the system produces a range of articulations and learns to model their acoustic consequences; 2) *comprehension*, in which acoustic input from an adult is mapped via phonology onto a semantic representation; 3) *imitation*, in which the system generates articulations in an attempt to reproduce the acoustic and phonological representations it previously derived from an adult utterance; and 4) *intentional naming*, in which the system uses a semantic representation (perhaps derived from vision) to generate articulations via phonology, and then compares this semantic representation to the one produced by comprehension of the system's own utterance.

In the remainder of this paper, we develop a computational simulation of the framework which, although simplified, serves to establish the viability of the approach. We discuss the implications of the framework for a variety of empirical findings on phonological development in the subsequent General Discussion.

Simulation

Given the considerable scope of the framework depicted in Figure 1, a number of simplifications were necessary to keep the computational demands of the simulation within the limits imposed by available computational resources. The most fundamental of these was that, instead of using continuous time and a fully recurrent network (Pearlmutter, 1989), the simulation used discrete time and a simple recurrent network (hereafter SRN; Elman, 1990).[2] This change results in a drastic reduction in the computational demands of the simulation because, once the states of context units are set, processing in an SRN is entirely feedforward. The drawback of this simplification is, of course, that we cannot capture the true temporal complexities of articulatory and acoustic representations, nor their interactions with phonological and semantic representations. In addition, variability due to speaker pitch and rate was not included—these are issues to be addressed by future work.

In order to reformulate the general framework in Figure 1 as an SRN, certain architectural changes are necessary. Specifically, in the framework, the mappings between phonology and semantics are assumed to be bidirectional and interactive. Given that processing in an SRN must be feedforward, the *input* mapping (from phonology to semantics) and the *output* mapping (from semantics to phonology) must be separated and implemented with separate units and connections. The resulting architecture (ignoring hidden and context units) maps output semantics $\rightarrow$ output phonology $\rightarrow$ articulation $\Rightarrow$ acoustics $\rightarrow$ input phonology $\rightarrow$ input semantics (where "$\Rightarrow$" corresponds to the forward model). It is important to keep in mind, however, that the division of input and output representations for phonology and semantics is not intended to imply that these representations are actually separate in child and adult speakers—to the contrary, we claim that the same representations underly both comprehension and production. Accordingly, the functional correspondence of the SRN architecture to the more general framework is maintained by running only subsections of the network and by copying unit states and targets as appropriate. This ensures that, for both semantics and phonology, the representations on the input and output sides of the network are identical.

Finally, the implementation does not include a physical articulatory apparatus that produces real sound, nor an auditory transducer that generates acoustic inputs from sound. Rather, these physical processes were approximated by coupled equations that map any combination of values over a set of articulatory variables onto a particular set of values over a set of acoustic variables (see below). These values are what the network's articulatory and acoustic representations encode. Considerable effort was spent to make these equations as realistic as possible while staying within the constraints of computational efficiency.

Despite these simplifications, the implemented simulation nevertheless embodies a solution to the fundamental computational challenges of speech comprehension and production discussed above. To the extent that it succeeds in learning to understand and produce words, it provides support for the more general framework and approach to phonological development.

Network Architecture

Figure 2 shows the fully recurrent version of the general framework, and the specific architecture of the implemented simple recurrent network version. The recurrent version (Figure 2a) is equivalent to Figure 1 with the addition of the groups of hidden units and connections that carry out the mappings among acoustics, semantics, and articulation. The implemented version (Figure 2b) has a general feedforward structure, starting from the upper right of the network and continuing clockwise to the upper left. In addition, as an SRN, the states of particular groups of units on the previous time step are copied to additional *context* units

[2] An SRN differs from a fully recurrent network primarily in that performance error is attributed to activations only for the current and immediately previous time step, but not further back in time (see Williams & Peng, 1990). Also, in an SRN, computing the output for a given input involves a single pass through the network whereas, in a fully recurrent network, it typically involves multiple iterations as the network settles to a stable state.

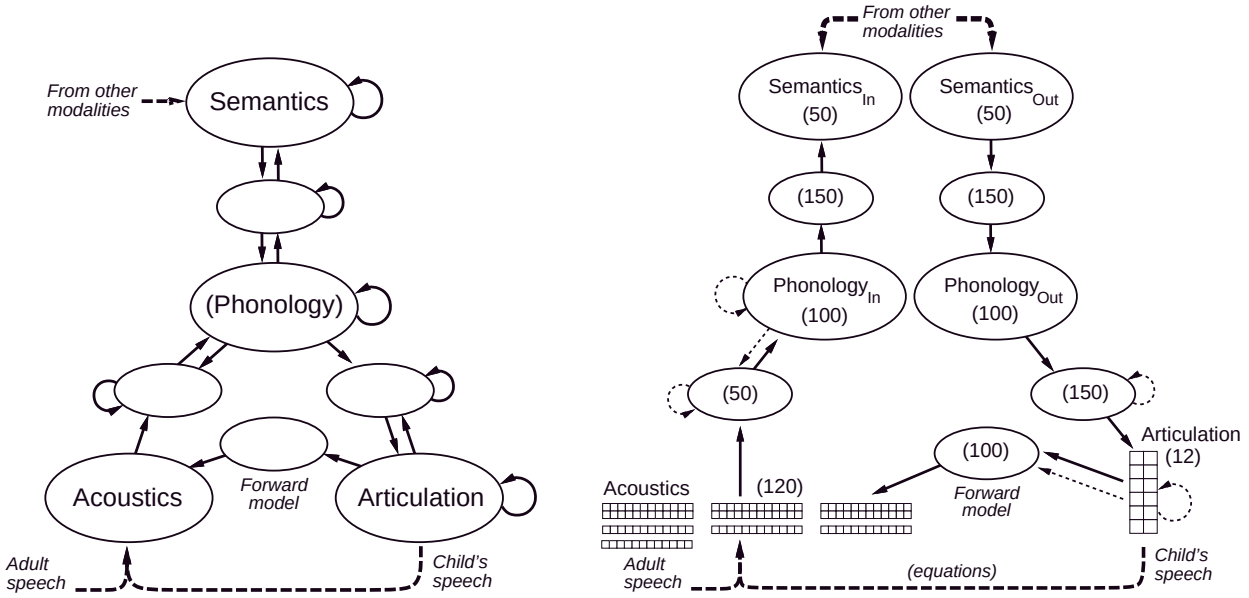

Figure 2. The architectures of (a) a fully recurrent version of the general framework in Figure 1, and (b) its instantiation as a simple recurrent network (SRN). Ovals represent different groups of units (as do the groups of small squares for Acoustics and Articulation); the number of units in each group is indicated in parentheses. Each solid arrow represents a standard projection of full connectivity from one group of units to another. The thin dashed arrows in (b) represent projections from "context" units whose states are copied from the previous time step; for the sake of clarity, the projections are depicted as coming from the source of the copied states rather than from separate context units (which are not shown). The thick dashed arrows represent external input: semantic patterns from other modalities (targets for comprehension; inputs for intentional naming), acoustic input sequences from an adult, and the acoustics generated by the system's own articulations.

and, thus, are available to influence processing on the current time step. It is by virtue of connections from these context units that the network can learn to be sensitive to non-local temporal dependencies in the task (see Elman, 1990, for discussion). Note that the actual context units are not shown in the figure; rather, the connections from context units are depicted as recurrent or feedback projections (shown as thin dashed arrows) from the source of the copied activations. Also note that there is one exception to this: In addition to a standard projection (solid arrow) from Articulation to the hidden units within the forward model, there is also a feedforward "context" projection (dashed arrow). These two arrows are intended to indicate that the patterns of activity over Articulation from both the current and previous time steps are provided as input to the forward model. Functionally equivalent groups that were split for the purposes of implementation as an SRN, but would be a single group within a fully recurrent implementation, are named with the subscripts *in* and *out* for clarity (e.g., Phonology$_{in}$, Semantics$_{out}$).

During comprehension and production, only Acoustics and Semantics$_{out}$ receive external input, and only Semantics$_{in}$ has externally specified targets. All of the remaining groups in the network, with the exception of Articulation, are "hidden" in the sense that their representations are not determined directly by the training environment but develop under the pressure of performing the relevant tasks accurately. Articulation has an unusual status in this regard. It has a predefined representation in the sense that unit activations have fixed and externally specified consequences for the resulting acoustics, and it is driven externally during babbling (see below). On the other hand, the network itself is given no information about

the implications of this representation; it must learn to generate the appropriate activations based solely on observing the consequences of these activations (much like actions on a musical instrument have predefined acoustic consequences, but a musician must learn which combination of actions produce a particular piece of music).

Corpus

The network was trained on the 400 highest (verbal) frequency monosyllabic nouns and verbs in the Brown corpus (Kučera & Francis, 1967) with at most four phonemes (mean = 3.42). Words were selected for presentation during training in proportion to a logarithmic function of their frequency of occurrence. As the main goal of the current simulation was to establish the viability of the general approach, little effort was made to structure the vocabulary of the model to correspond to the actual language experience of infants and young children. The restriction to monosyllables was for reasons of simplicity and because monosyllables dominate the production of children acquiring English (Boysson-Bardies, Vihman, Roug-Hellichius, Durand, Landberg, & Arao, 1992; Vihman, 1993); note, however, there there is nothing in the current implementation that precludes applying it to multisyllabic or even multiword utterances.

Representations

The following descriptions provide a general characterization of the representations used in the simulation. Many details have been omitted for clarity, particularly with respect to the articulatory and acoustic representations and the mapping between them—see Plaut and Kello (in preparation) for full details and equations.

Articulatory Representations. The articulatory domain was carved into discrete events, roughly at points of significant change in articulation.[3] Each event was represented with six articulatory degrees of freedom based on configurations of oral/facial muscles that are more-or-less directly relevant to speech. Each dimension was a real value in the range $[-1,1]$, corresponding to a static state in time. Three constraints entered into our choice of dimensions: 1) they needed to capture the physical similarities and differences in producing different phonemes in English, as well as the variations in producing the same phoneme in different contexts; 2) they had to lend themselves to engineering the equations that map articulation to acoustics; and 3) they had to be mostly independent of each other, in terms of muscular control, to avoid complications of building dependencies into the network architecture.

The six articulatory degrees of freedom are as follows (the labels are used in the example representations shown in Figure 3 below):

1. *Oral Constriction (ORC)*. This corresponds to the maximal amount of air flow for a given articulatory state. It represents the combined effects of constricting articulators such as the jaw and parts of the tongue. In general, vowels have low oral constriction, approximants have moderate constriction, fricatives have high constriction, and stops have complete constriction.
2. *Nasal Constriction (NAC)*. Since nasal constriction is primarily controlled by raising and lowering the soft palate (Ladefoged, 1993), this dimension directly corresponds to the height of the soft palate.
3. *Place of Constriction (POC)*. This corresponds to the location of the maximal point of constriction for a given articulatory state. Location is coded from front to back, with the most front value equal to a labial POC, and the most back value equal to a glottal POC. With moderate to high amounts of oral constriction, POC codes place of articulation for consonant-like articulations. With little oral constriction, articulation becomes vowel-like and POC has no effect on acoustics.

[3] As pointed out above, discretization of time was a simplification introduced to permit the the use of an SRN architecture. We believe that continuous dynamics would more accurately capture a number of relevant phenomena.

4. *Tongue Height (HEI)*. This roughly codes the maximal closeness of the tongue to the roof of the mouth, but only when there is little oral constriction. Tongue height is directly related to the openness of a vowel.
5. *Tongue Backness (BAK)*. This is directly related to vowel backness. As with tongue height, it only takes effect when there is little oral constriction.
6. *Voicing (VOC)*. This corresponds to a combination of glottal opening and vocal chord constriction, given that our model assumes constant pulmonary pressure for the sake of simplicity.

Each articulatory degree of freedom in the network was coded by two units whose activities fell in the range [0,1]. The value for each degree of freedom was coded by the signed difference between these values, having a possible range of $[-1,1]$. Note that the same value can be coded by different combinations of unit activities; the network can learn to use any of these. Finally, the network represented both the past and current articulatory event in order to capture articulatory change information that has consequences for acoustics.

Acoustic Representations. As with articulation, the acoustic domain was divided into discrete events. We chose ten "acoustic" dimensions based on aspects of the speech environment that are generally thought to be relevant to speech perception. The choice of dimensions was constrained by the same factors as the articulatory dimensions (see above). The word acoustic is quoted above because, although sound is certainly the primary perceptual/proprioceptive domain of speech, oral dimensions such as jaw openness also play a role. For example, newborns pay special attention to certain mouth posturings (Meltzoff & Moore, 1977), and three- to four-month-old infants are aware of the relation between certain facial and vocal activities (see Locke, 1995, for review). We incorporated the role of visual perception/proprioception by including a visual dimension of jaw openness in our acoustic representation. In fact, our model provides a reason for why infants might pay such close attention to the mouth: Visual speech helps to reduce the motor equivalence problem, thus facilitating the acquisition of both the comprehension and production of speech.

Another important aspect of our acoustic representation is that the dimensions were normalized for overall speaker variations such as rate and pitch (but random variation and coarticulation were incorporated; see below). Researchers have shown that infants as young as two months can normalize for speaker variation (Kuhl, 1983; Jusczyk, Pisoni, & Mullennix, 1992). Although this simplification seems reasonable, we believe that the mechanism for speaker variability is an important part of understanding phonological development, and we intend to incorporate this into future versions of the model.

Unlike the articulatory degrees of freedom, each acoustic dimension had an accompanying *amplitude* value corresponding to the degree to which information on that dimension was present in a given acoustic event. For example, first formant frequency was an acoustic dimension, yet not all acoustic signals have formant structure (e.g., a voiceless fricative such as /s/). In this case, the degrees of freedom corresponding to formants would be assigned very low amplitudes in our encoding.

The ten acoustic dimensions are as follows (note that we have included here some aspects of the articulation-to-acoustics mapping; further detail is provided below):

1–3. *First through third formant frequencies.* Since these are normalized, they essentially code frequency position relative to a particular formant and a particular speaker. In general, the amount of periodicity in the putative acoustic wave form determines the amplitudes of these dimensions.

4–6. *First through third formant transitions or derivatives.* These code the normalized amount of rise or fall in each format from the previous to the current acoustic state. Their amplitudes are identical to those for the corresponding formant frequency.

7. *Frication.* This is a composite of the frequency, spread, and intensity of very high frequency, non-periodic energy. Previous research suggests that all of these measures play a role in distinguishing different fricative sounds, although the exact relationships are unclear (see Lieberman &

Blumstein, 1988, for review). We simplified matters by collapsing these variables into one acoustic dimension. Frication is present when oral and/or nasal constriction is slightly less than fully constricted.

8. *Burst.* This dimension is a simplification analogous to frication: a composite of acoustic measures that are involved in distinguishing plosive releases of different places of articulation. Bursting is present when a sufficiently large positive change occurs in oral or nasal constriction.
9. *Loudness.* This codes the overall amount of acoustic energy in a given event (normalized for speaker and stress). For example, silence has very little or no energy, fricatives have little energy, nasals have significantly more, and vowels have the most. The amplitude value for loudness is redundant and therefore permanently set to one.
10. *Jaw openness.* This dimension reflects the size of the mouth opening for a given articulatory state, and was computed based on the amount and place of oral constriction. We chose jaw openness because it is relatively visible (and presumably salient), but since a speaker cannot normally see their own jaw, this dimension also represents the analogous proprioceptive signal. Since the jaw is always somewhere along the continuum of openness, the amplitude value was unnecessary and therefore set to one.

Each of the ten acoustic dimensions were represented in the network with a separate bank of twelve units, laid out in one dimension corresponding to the $[-1,1]$ range of possible dimension values. A specific acoustic value was encoded by Gaussian unit activity with mean equal to the value and fixed variance. Unit activities were determined by sampling this Gaussian at the 12 equal-spaced intervals indicated by the unit positions within the bank. The total activity (i.e., the area under the Gaussian) was set to equal the amplitude of that acoustic dimension. Whereas articulation was represented by both a past and current articulatory vector in order to capture change information, an acoustic event contains dimensions that directly code change in time. Therefore, the network represented a single acoustic event that changed over time.

Articulation-to-Acoustics Mapping. The translation from two articulatory events (past and current) to an acoustic event consisted of ten equations, one for each acoustic dimensions. Each equation was written solely in terms of one or more of the articulatory degrees of freedom. The functions ranged from primarily linear and consisting of one or two variables, to highly non-linear and consisting of four or five variables. We will not present the details of the equations here (see Plaut & Kello, in preparation) but we outline the major relationships they embodied.

1. Formant frequency is determined primarily by either tongue height and backness, or by POC. In addition, nasal constriction tends to lower the first formant frequency (Lieberman & Blumstein, 1988).
2. Formant transitions are based on change in oral and/or nasal constriction, in combination with change in the dimensions relevant to formant frequencies. The relationship between place of articulation in a plosive release and the following vowel played a major role in determining these equations (see Liberman, 1996).
3. The amplitudes of the six formant dimensions are based on a combination of voicing and both oral and nasal constriction.
4. The frication and burst values are determined by place of articulation. The amplitudes of these dimensions are based on current (and, for bursting, also previous) oral and nasal constriction values.
5. Loudness is determined by a combination of all six articulatory degrees of freedom.

To illustrate the articulatory and acoustic representations and the mapping between them, Figure 3 shows the sequence of events corresponding to an "adult" utterance of the word SPIN /spɪn/ (see the "Adult Utterances" subsection below for a description of how such sequences are generated).

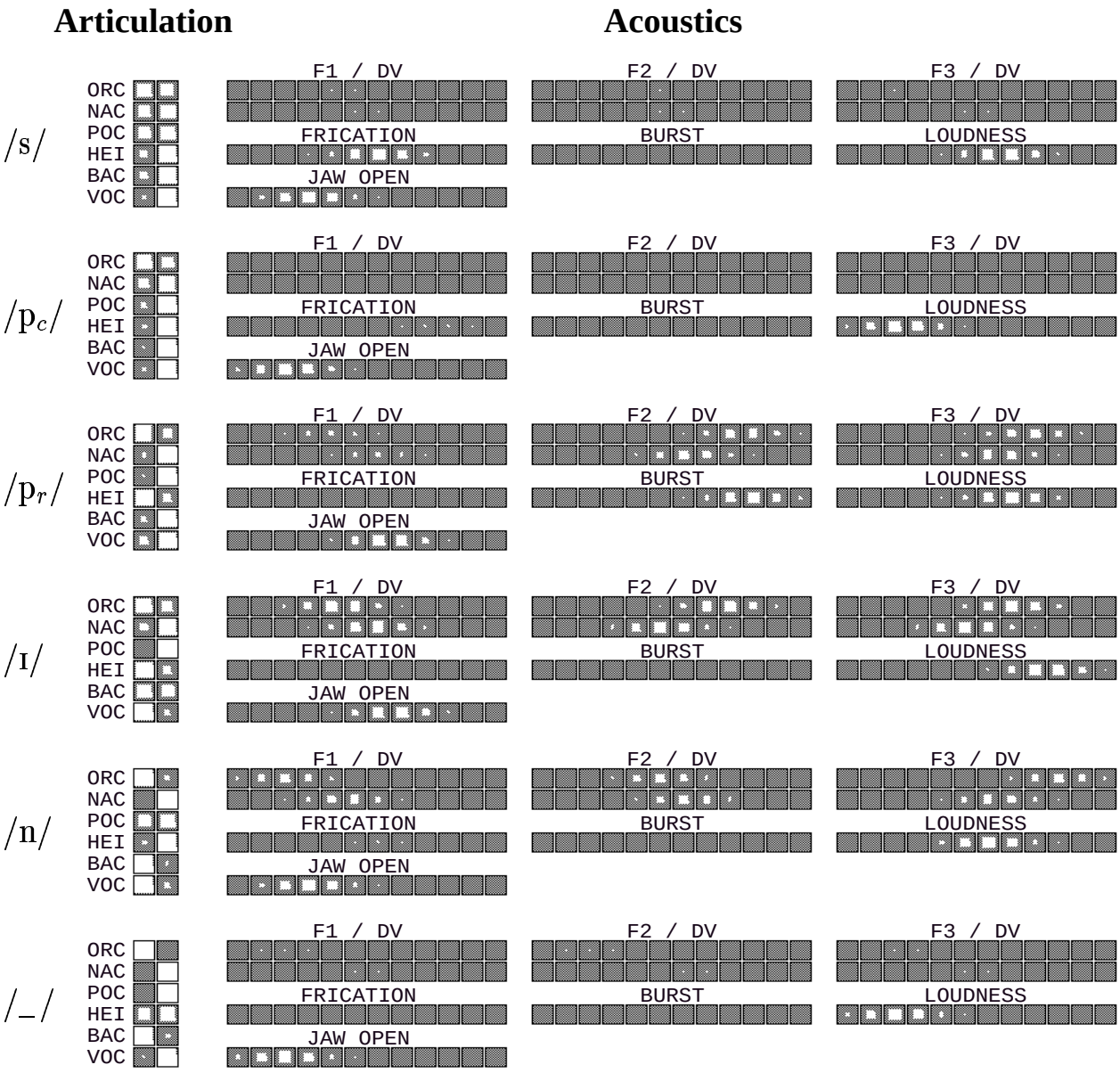

Figure 3. The sequence of articulatory (left) and acoustic (right) events corresponding to an adult utterance of the word SPIN. In the simulation, articulatory degrees of freedom are subject to intrinsic variability within permissible ranges; for illustration purposes, this variability was not applied to the depicted events. These events were, however, still subject to both perseverative and anticipatory coarticulation. Each acoustic event is determined by both the current and previous articulatory event (which, for the first, is a copy of the current event). See the text for definitions of each unit group.

Babbling. In our theoretical framework, babbling serves to train a forward model to learn the relationship between articulation and acoustics. The system uses this model to convert acoustic error to articulatory error to incrementally improve its own productions during imitation and intentional naming. A given babbling episode consisted of a string of articulations generated stochastically from a model of babbling behavior that combined canonical and variegated babbling. This model was intended to approximate the articulatory effects of a bias towards mandibular (jaw) oscillation in the oral behavior of young infants. This oscillatory bias has been proposed as a fundamental constraint on early speech development (Davis & MacNeilage, 1995; Kent, Mitchell, & Sancier, 1991; MacNeilage, in press; MacNeilage & Davis, 1990) and is consistent with the central role of rhythmic behavior in motor development more generally (Thelen, 1981; Thelen & Smith, 1994). The generated articulatory sequences reflected both reduplicated babbling (to the extent that POC remained constant across the articulatory string) and variegated babbling (to the extent that POC changed).

Specifically, each instance of babbling was composed of five articulatory events. Four of the six degrees of freedom in the first event of a sequence were given random values; the remaining two—oral and nasal constriction—-were random but their probabilities were weighted towards the endpoints (i.e., -1 and 1, which correspond to completely closed and opened, respectively). Each subsequent event in a sequence was generated by sampling from a Gaussian probability distribution centered around the previous values for 5 of the 6 degrees of freedom (producing a bias towards gradual change); oral constriction was stochastic but weighted towards the opposing sign of the previous value, thus exhibiting a bias towards oscillation. The resulting articulatory sequence constituted the babbling utterances used to train the articulatory-acoustic forward model, as described below.

Adult Utterances. Adult word utterances were derived from canonical strings of phonemes, each ending with a null or silent phoneme. Note that, although the canonical forms were represented as phonemes, the resulting acoustic events did not have a 1-to-1 correspondence with phonemes (see below). Moreover, each articulatory event underlying an acoustic pattern was subject to considerable intrinsic variability and was shaped by multiple phonemes due to coarticulatory influences.

The procedure for generating an instance of an adult utterance of a word was as follows. The word's phoneme string was first converted into a sequence of articulatory events. Most phonemes corresponded to a single articulatory event, although plosives and dipthongs were coded as two events (except that the second, "release" event of the first of two adjacent plosives was deleted). The articulatory event(s) corresponding to a phoneme were defined in terms a range of permissible values over the articulatory degrees of freedom. The center and size of each range was estimated from the role and importance that a given degree of freedom plays in producing a given phoneme (based on phonetic research drawn largely from Ladefoged, 1993). A specific utterance was generated by randomly sampling from a Gaussian distribution centered on the ranges for each phoneme, and clipped at the endpoints. The randomly determined initial values for each articulatory event were then adjusted based on both perseverative and anticipatory coarticulation. For a given event, each degree of freedom was first pulled towards the specific values of previous events, with the strength of this pull scaled by the number of intervening events and the value range of the perseverative influence. Then, each degree of freedom was pulled towards the canonical values of the subsequent events (i.e., the range midpoints, before their values were randomly generated), scaled by the same factors. Finally, the coarticulated values were forced to be within their respective ranges. The resulting string of articulatory events was input to the articulation-to-acoustics equations to generate a string of acoustic events. In generating the first acoustic event, the first articulatory event was interpreted as both the current and previous articulatory state, analogous to allowing for articulatory preparation before beginning an utterance. Note that, although the articulation-to-acoustics equations are deterministic, adult utterances of a given word exhibited considerable variability due to the random sampling of articulatory degrees of freedom within the permissible range for each articulatory event. Moreover, individual tokens of the same phoneme varied both from this random sampling and from the coarticulatory influences of surrounding phonemes.

Semantic Representations. No attempt was made to design semantic representations that capture the actual meanings of the 400 words in the training corpus. Rather, artificial semantic representations were developed which, while constituting only a very coarse approximation to the richness and variability of actual word meaning, nonetheless embodied the core assumptions behind the challenges of comprehension and production: namely, that semantic similarity is unrelated to acoustic/articulatory similarity.

Specifically, the semantic representations of words were generated to cluster into artificial semantic "categories" (Chauvin, 1988; Plaut, 1995, 1997). Twenty different random binary patterns were generated over 50 semantic features, in which each feature had a probability p_a = .2 of being active in each prototype. Twenty exemplars were then generated from each prototype pattern by randomly altering some of its features; specifically, each feature had a probability of .2 of being resampled with p_a = .2. The effect of this manipulation is to make all exemplars within a category cluster around the prototype, and for all semantic patterns to have an average of 10 active features (range 6–16) out of a total of 50. The resulting 400 semantic patterns were then assigned to words randomly to ensure that the mappings from acoustics to semantics and from semantics to articulation were unsystematic.

Training Procedure

As mentioned in the Introduction, the system undergoes four types of training experiences: babbling, comprehension, imitation, and intentional naming. Although each of these is described separately below, it is important to keep in mind that they are fully interleaved during training.

Babbling. The role of babbling in our framework is to train an articulatory-acoustic forward model; the only part of the network involved in this process is Articulation, Acoustics, and the hidden units between them (see Figure 2). First, a particular sequence of articulations corresponding to an instance of babbling was generated (as described under "Representations" above). This sequence was then passed through the articulation-to-acoustics equations to produce a sequence of "actual" acoustic patterns. The articulations also served as input to the forward model (see Figure 2), which generated a sequence of predicted acoustic patterns. The discrepancy or error between the actual and predicted acoustics at each step, measured by the *cross-entropy*[4] between the two patterns (Hinton, 1989), was then back-propagated through the forward model and used to adjust its connection weights to improve its ability to predict the acoustic outcome of the given articulations. In this way, the forward model gradually learned to mimic the physical mapping from articulatory sequences to acoustic sequences (as instantiated by the articulation-to-acoustics equations).

Comprehension. Comprehension involves deriving the semantic representation of a word from a sequence of acoustic patterns corresponding to an "adult" utterance of the word (generated as described under "Representations" above). Given such an utterance, the network was trained to derive the semantics of the word in the following way. Prior to the first acoustic event, the context units for Phonology$_{in}$ and for the hidden units between Acoustics and Phonology$_{in}$ (see Figure 2) were initialized to states of 0.2. (Note that there are no context units between Phonology$_{in}$ and Semantics$_{in}$—this means that, although the pattern of activity over phonology changes as acoustic input comes in, this activity must map to semantics in parallel.) Then, each acoustic event was presented successively over Acoustics and the activations of all units between Acoustics and Semantics$_{in}$ were computed. For each event, the resulting semantic pattern was compared with the correct semantics for the word and the resulting error was propagated back through the network to Acoustics to accumulate weight derivatives for connections in this portion of the network (including the connections from context units). After processing each event, the activities of the Phonology$_{in}$ units and the Acoustics-to-Phonology$_{in}$ hidden units were copied to their corresponding context units, to influence processing of the next acoustic event. After the last acoustic event was presented (corresponding to silence), the accumulated weight derivatives were used to adjust the weights to improve comprehension performance

[4]The cross-entropy between a pattern of activation over a set of units and their target activations is given by $-\sum_i t_i \log_2(a_i) + (1-t_i)\log_2(1-a_i)$, where a_i is the activity of unit i and t_i is its target.

on subsequent presentations of the word. Because error was based on the discrepancy of the generated and correct semantics from the very beginning of the acoustic input sequence, the network was pressured to derive information about the semantics of the incoming word as quickly as possible.

Imitation. Imitation involves using a phonological representation derived from an adult utterance as input to drive articulation, and comparing the resulting acoustic and phonological representations with those of the adult utterance. Imitation could, in general, be based on any adult utterance, including those without any clear semantic referent. However, for practical reasons (i.e., to avoid having a much larger training corpus for imitation than for comprehension), training episodes of imitation were yoked to episodes of comprehension in the simulation.

Specifically, after having processed an adult utterance of a word for comprehension (see above), the final phonological pattern over Phonology$_{in}$ was copied to Phonology$_{out}$ (see Figure 2; recall that these correspond to the same group in the general framework). The phonological pattern then served as the static input "plan" for generating an articulatory sequence. All of the context units between Phonology$_{out}$ and Phonology$_{in}$ were initialized to states of 0.2. The network then computed unit activations up to and including Articulation. This articulatory pattern and the one for the previous step (which, for the first step, was identical to the current pattern) were mapped through the forward model to generate predicted acoustics. At the same time, the articulatory patterns were used to generate an actual acoustic event via the articulation-to-acoustics equations. The discrepancies between the network's actual acoustics and those predicted by the forward model were used to adapt the forward model, as during babbling. In addition, the actual acoustics were mapped by the comprehension side of the network to generate a pattern of activity over Phonology$_{in}$. The error between the activations generated by the network and those generated by the adult were then calculated both at Phonology$_{in}$ and at Acoustics.[5] The phonological error was back-propagated to acoustics (without incrementing weight derivatives) and added to the acoustic error. The combined error was then back-propagated through the forward model (again without incrementing derivatives) to calculate error for the current articulatory pattern. This error was then back-propagated to Phonology$_{out}$ and weight derivatives were accumulated for the production side of the network. At this point, the relevant unit activations were copied onto context units, and the pattern over Phonology$_{out}$ was used to generate the next articulatory event. This process repeated until the network produced as many articulatory events as there were acoustic events in the imitated adult utterance (n.b. a more general strategy would be to continue articulating until a number of silent acoustic events are produced).

Intentional Naming. The final type of training experience included in our general approach to phonological development is the intentional generation of an utterance on the basis of semantic input (perhaps derived from another modality). In the current simulation, however, the ability was trained only indirectly.

Again, for reasons of efficiency, intentional naming was yoked to comprehension. After the network was trained to comprehend a given word, the correct semantic pattern for the word was then presented as an input pattern over Semantics$_{out}$ (this pattern had been assumed to be available as targets over Semantics$_{in}$). This input pattern was then mapped by the network to generate an output pattern over Phonology$_{out}$. The error between this pattern and the one over Phonology$_{in}$ coding the entire adult utterance (again, these should be thought of as the same group) was then back-propagated to Semantics$_{out}$, and the corresponding weights were adapted to improve the network's ability to approximate its own phonological representation of the adult's utterance given the corresponding semantics as input. Note that the targets for the task change as the network's own phonological representations evolve during the development of comprehension; the output side of the system must nonetheless track this change. Eventually, both the comprehension and production system converge on being able to map the same phonological representation both to and from semantics. In

[5]Note that the use of a direct comparison between the acoustics generated by the adult and those generated by the network assumes a considerable amount of pre-acoustic normalization. It should be noted, however, that training imitation using comparisons only at the phonological level results in only marginally worse performance (see Plaut & Kello, in preparation).

fact, this training procedure was intended to approximate the effects of bidirectional interactions between phonology and semantics in a fully recurrent version of the system.

Once the system has the capability of mapping a semantic representation onto a phonological representation, training during imitation enables this phonological pattern to be mapped to an analogous one on the input side, which can then be mapped to the corresponding semantics due to training during comprehension. In this way, the network can be tested for its ability to map a semantic pattern via phonology to an articulatory sequence which, according to its own comprehension system, sounds the same and means the same as the intended utterance. The entire mapping (from Semantics$_{out}$ to Semantics$_{in}$) was not, however, explicitly trained because the resulting back-propagated error derivatives would be very small (due to the large number of intervening layers of units) and because children seem relatively insensitive to the semantic plausibility of their own utterances (e.g., the *fis* phenomenon; Berko & Brown, 1960; Dodd, 1975; Smith, 1973).

Testing Procedure

The network was trained on 3.5 million word presentations and babbling episodes. Although this may seem like an excessive amount of training, children speak up to 14,000 words per day (Wagner, 1985), or over 5 million words per year. For each word presentation, the network was trained to comprehend the word and then to imitate the resulting acoustics and phonology. The network also produced and learned from an unrelated babbling sequence. Although, in actuality, the onset of babbling precedes clear attempts at imitation, and falls off as production skill increases, the model engages in both babbling and imitation throughout training. Babbling and early word production do, in fact, overlap in time to a large degree (see Vihman, 1996), and the phonetic inventory that children use when beginning to produce words is drawn largely from the inventory of sounds produced during babbling (Vihman & Miller, 1988). Moreover, some babble-like utterances may result from undeveloped speech skills during attempts to imitate or intentionally produce words. Such attempts nontheless constitute opportunities for learning the relationship between articulation and acoustics; as mentioned above, our network adapts its forward model during both babbling and imitation. As it turns out, the performance of the forward model achieves a reasonable level of accuracy fairly quickly, so continuing to include babbling episodes throughout training has little impact on performance.

After every 500,000 word presentations during training, the network was evaluated for its ability to comprehend and imitate adult speech, and to produce comprehensible intentional utterances. Because there was intrinsic variability among adult utterances, the network was tested on 20 instances of each of the 400 words in the training corpus. Performance on each task was based on whether each stimulus could be accurately discriminated from all other known words, using a *best-match* criterion over semantics and, for imitation, also over phonology. Specifically, when applied to semantics, the network was considered to comprehend a word correctly if its generated semantics matched the correct semantics for that word better (in terms of normalized dot product) than the semantics for any other word in the training corpus. The best-match procedure over phonology was a bit more complicated as there are no predefined phonological representations for words against which to compare the network's utterance. Moreover, due to intrinsic variability, different adult instances of a word generated somewhat different phonological representations during comprehension. Accordingly, the best-match criterion was applied to phonology by comparing the phonological pattern generated by the network's utterance of a word with the phonological representations generated by all 8000 adult utterances (20 instances of 400 words) and considering the network's utterance correct if the best-matching adult utterance was one of the 20 instances of the word.

Much of the most important evidence on the nature of phonological development comes from an analysis of children's speech errors (Ingram, 1976; Menn, 1983; Smith, 1973; see Bernhardt & Stemberger, 1997, for review). Although a comprehensive account of the systematicity and variability of child speech errors remains a long-term goal of the current approach, an initial attempt can be made by examining the nature of

the network's errors in comprehension, imitation, and intentional naming. Specifically, if the best match to the network's utterance was the representation of a word other than the stimulus, the network was considered to have made an error, which was then evaluated for phonological and/or semantic similarity with the stimulus.[6] For these purposes, an error was considered phonologically related if it differed from the stimulus by an addition, deletion, or substitution of a single phoneme, and it was considered semantically related if its assigned semantic pattern came from the same artificial category (see "Representations" above).

Results

Due to limitations on space, we present only a subset of the results derived from the network (see Plaut & Kello, in preparation, for a more comprehensive presentation). In particular, we omit a full characterization the performance of the forward model. This model is acquired fairly rapidly, achieving reasonable performance within the first 1 million word presentations and babbling episodes. Moreover, its ability to predict the acoustic consequences of articulatory events is ultimately very accurate, producing an an average cross-entropy per event of less than 0.39 summed over the 120 Acoustic units (cf. 33.7 at the beginning of training). When inaccurate, it tends to produce Gaussian activity over the acoustic banks with greater variance than in the actual acoustic signal, which is a natural indication of reduced confidence in the underlying dimension value.

Correct Performance. Figure 4 shows the correct performance of the network over the course of training on comprehension, imitation, and intentional naming using the best-match criterion. First note that comprehension performance improved relatively rapidly, reaching 84.3% correct by 1M (million) word presentations and 99.6% by 3.5M presentations. This level of performance is impressive given the lack of systematicity in the mapping between acoustics and semantics and the considerable intrinsic variability of adult utterances. Relative to comprehension, competence in production developed more slowly: When evaluated at semantics, the network was only 54.2% correct at imitation by 1M presentations, although it did achieve 91.7% correct by 3.5M presentations. Intentional naming was slightly poorer than imitation throughout training, eventually reaching 89.0% correct. This is not surprising as the task involves mapping through the entire network and was not trained explicitly.

When evaluated at phonology, imitation performance was more accurate, achieving 96.5% correct by 3.5M word presentations. The rapid rise in best-match imitation performance at phonology at 0.5M presentations is due to the fact that phonological representations are relatively undifferentiated at this point.

Overall, the network achieved quite good performance at both comprehension and production. The fact that comprehension precedes production in the model stems directly from the fact that learning within the production system is driven by comparisons over representations within the comprehension system. The excellent performance on imitation, particularly at phonology, demonstrates that feedback from the comprehension system via a learned forward model can provide effective guidance for articulatory development. The findings provide encouraging support for the viability of our general framework for phonological development.

Error Analysis. To further characterize the network's performance, we analyzed the errors made by the network under the best-match criterion after 3.5M word presentations. As described above, an error response was considered semantically related to the correct (target) word if it belonged to the same category, and phonologically related if it differed from the target word by the addition, deletion, or substitution of a single phoneme. Based on these definitions, errors were classified as *semantic* (but not phonological), *phonological* (but not semantic), *mixed* semantic and phonological, or *miscellaneous* (neither semantic nor

[6] This way of defining speech errors allows for only word responses, which is clearly inadequate. Determining nonword error responses is problematic, however, because the acoustic sequences generated by the network cannot be interpreted directly. Plaut and Kello (in preparation) address this problem by training a separate network to map sequences of acoustic events onto sequences of phonemes, analogous to a trained linguistic listening to and transcribing speech.

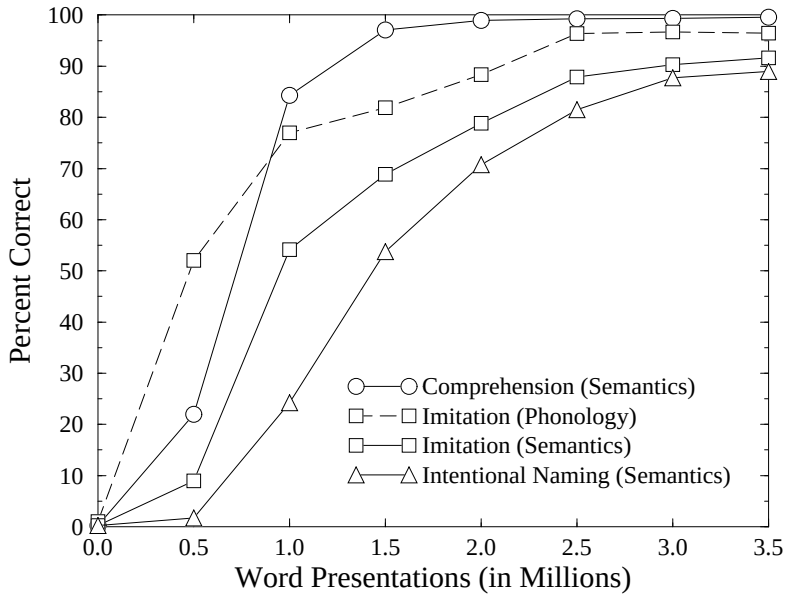

Figure 4. Correct performance of the network in terms of a best-match criterion, on comprehension (at semantics), imitation (at both phonology and semantics), and intentional naming (at semantics), over the course of training.

phonological, although many such errors exhibited phonological similarity that failed to satisfy our strict definition).

Table 1 presents the percentage of error occurrences for each of the above types and for each task, as well as the chance rates of occurrences of each error type. Calculating the chance probability of a semantic error was straightforward as there were five equiprobable categories. The chance probability for a phonological error was determined empirically by computing the rate of phonological relatedness among all pairs of trained words. The chance rate of a mixed error was simply the product of these two probabilities (i.e., we assumed independence, given that semantic patterns were assigned to words as phoneme strings randomly).

In general, the network showed a strong bias toward phonological similarity in its errors compared with the chance rate, for both comprehension and imitation (although this is based on relatively few errors in the former case). Comprehension and imitation also produced semantic errors below chance, whereas the rates for intentional naming were well above chance. Interestingly, the rate of mixed errors in imitation, although low, was almost five times the chance rate, consistent with a general preference in production for errors sharing both semantic and phonological similarity with the target (see, e.g., Levelt, 1989).

Although comprehension errors were rare, when they occurred they typically involved the addition of a final plosive (most commonly /t/ or /d/) to yield a higher frequency word (e.g., PASS /pæs/ → PAST /pæst/), presumably because utterance-final silence was misinterpreted as the closure event of the plosive, and the system is insufficiently sensitive to the acoustics of the release. This type of error was also common in imitation, presumably because its feedback during training was derived from comprehension. Imitation also produced a number of multi-change error which appear to be phonetically conditioned (e.g., GAVE /geɪv/ → CAME /keɪm/), and some evidence of cluster reduction (e.g., FLAT /flæt/ → MATCH /mætʃ/), although the corpus provided very few opportunities to observe the latter.

At the phoneme level, there were far more errors on consonants than on vowels and, among consonants, a relatively higher error rate on fricatives, affricates (e.g., /tʃ/) and /ŋ/ (as in RING). These errors involved both additions and deletions; when they were deleted, they were often replaced by a plosive. In fact, plosives accounted for over half of the total number of insertions. By contrast, the liquids /r/ and /l/ were deleted

Table 1
Error Rates and Percent of Error Types for Various Tasks, and Chance Rates

		Error Type			
Task	Error Rate	Semantic	Phonological	Mixed	Misc.
Comprehension	0.45	14.1 (0.71)	85.3 (32.8)	0.0 (0.00)	0.0 (0.00)
Imitation	3.52	10.1 (0.51)	74.3 (28.6)	1.9 (4.75)	13.7 (0.18)
Intentional Naming	11.0	40.9 (2.05)	11.4 (4.4)	0.0 (0.00)	47.7 (0.62)
Chance		20.0	2.6	0.4	77.0

Note: Each value in parentheses is the ratio of the observed error rate to the Chance rate listed at the bottom of the column. Performance was measured after training on 3.5 million word presentations. "Error Rate" is based on 20 instances of adult utterances of each of 400 words (8000 total observations) for comprehension and imitation, and one instance for intentional naming (because the network's own articulations are not subject to variability). Comprehension and intentional naming were evaluated at semantics, whereas imitation was evaluated at phonology.

occasionally, but never inserted. These characteristics are in broad agreement with the properties of early child speech errors (e.g. Ingram, 1976).

Although these error analyses are preliminary, and are significantly limited by allowing only word responses, they suggest that the current approach can be applied fruitfully to understanding the nature of children's speech errors.

General Discussion

Infants face a difficult challenge in learning to comprehend and produce speech. The speech stream is extended in time, highly variable, and (for monomorphemic words) bears little systematic relationship to its underlying meaning. Articulatory skill must develop without any direct instruction or feedback, and comprehension and production processes must ultimately be tightly coordinated.

The current paper outlines a general framework, based on principles of connectionist/parallel distributed processing, for understanding how the infant copes with these difficulties, and presents an implemented simulation which, although simplified relative to the framework, nonetheless instantiates its fundamental hypotheses. In particular, two key properties of the framework and implementation reflect a close interplay between comprehension and production in phonological development. The first is that both comprehension and production are subserved by the same underlying phonological representations. These representations are not predefined but emerge under the combined pressures of mediating among acoustic, semantic, and articulatory information. The second key property is that the necessary articulatory feedback for the production system is derived from the comprehension system. Specifically, proximal (articulatory) error is derived from the distal (acoustic and phonological) consequences of articulation via a learned articulatory-acoustic forward model (also see Jordan, 1996; Perkell et al., 1995). The simulation demonstrates that a model instantiating these properties can, in fact, learn to cope with time-varying, variable speech in comprehension, and use the resulting knowledge to guide production effectively in imitation and intentional naming. Moreover, its pattern of errors, although not matching child speech errors in detail, does show the appropriate general trends, suggesting that the approach may provide a computationally explicit basis for understanding the origin of such errors.

The bulk of the current paper has focussed on the nature of the computational problems posed by phono-

logical development, and on the formulation of a particular approach for solving these problems. At this stage in the work, relatively little attention has been directed at relating the simulation, and the more general framework, to specific empirical phenomena concerning phonological development in infants and young children. In the remainder of the paper, we consider three categories of empirical phenomena—the relationship between comprehension and production, the time course of speech acquisition, and the phonetic content of speech errors—and address briefly how our approach might account for some the more central findings in each. While many of these points are addressed directly by the existing framework and implementation, some of them constitute important directions for future research.

Relationship Between Comprehension and Production

One of the most basic findings in phonological development is that skilled word comprehension precedes skilled word production (Benedict, 1979; Reznick & Goldfield, 1992; Snyder, Bates, & Bretherton, 1981). The finding has prompted speculation that a certain skill level of comprehension is necessary for production skills to advance. Our model embodies this notion in terms of the maturity of phonological representations that map acoustic input onto semantics. The internal representations must begin to stabilize in comprehension before their input to the production system becomes useful for learning articulation. Our approach differs from Vihman's (1996) idea of an "articulatory filter," in which the articulatory system mediates which words the child can perceive, remember, and therefore produce. Our framework holds that the perception and comprehension of speech can develop somewhat independently of the articulatory system, although the mature system must eventually mediate both tasks with the same underlying phonological representations.

Another finding that points towards a link between development in comprehension and production is that the phonetic distribution of late babbling (i.e., by 10 months) is influenced by the ambient language (Boysson-Bardies et al., 1992). When beginning to imitate and speak from intention, a child's utterances, as well as those of our network, will tend to sound like babble (i.e., highly variable exploration of articulation) because the link between phonology and articulation has not yet developed. If the adult characterizes these early attempts at speech as babble, then indeed babbling will tend to have phonetic characteristics of the ambient language. Similarly, the phonetic characteristics of early word production overlap to a large degree with the characteristics of a given infant's babbling (Stoel-Gammon & Cooper, 1984; Vihman, Maken, Miller, Simmons, & Miller, 1985). Given that the instantiation of babbling in the current model is not influenced by the developing phonological representations, the model does not account for this finding, but we plan to address it in future extensions.

Time Course of Speech Acquisition

A second major area of investigation in phonological development relates to the order of acquisition of various speech production skills. We consider two findings to be of particular importance. The first is that infants often have a small repertoire of "protowords" late in the babbling phase, but prior to true word production (Menyuk & Menn, 1979; Stoel-Gammon & Cooper, 1984; Werner & Kaplan, 1984). Protowords are characterized by relatively stable patterns of vocalization that serve to communicate broad distinctions between situations (e.g., request an object versus request social interaction). As discussed in the Introduction, distributed networks have an inherent tendency to map similar input patterns onto similar output patterns; this bias is overcome only gradually in learning an unsystematic mapping. In the current context, this means that broadly similar semantic patterns will map initially onto similar phonological patterns. If the phonological patterns are similar enough, they will be heard as the same utterance (i.e., a protoword).

The second finding is an example of the ubiquitous phenomenon of U-shaped learning: As the child learns to produce more words, production of originally well-learned words often regresses (Vihman & Velleman, 1989). More generally, articulation increases in variability and decreases in phonetic accuracy as de-

velopment progresses, although this trend reverses as articulatory skills approach mature levels. This pattern is generally interpreted as indicating a shift from a whole-word system to a more systematic, segmentally-based one (see Jusczyk, 1997; Vihman, 1996). A number of researchers (e.g., Jusczyk, 1986; Lindblom, 1992; Studdert-Kennedy, 1987; Walley, 1993) have pointed to the growth of receptive vocabulary as exerting a powerful influence on the degree to which phonological representations become segmental. Early on, relatively undifferentiated, "wholistic" phonological representations may suffice for discriminating among the few words known to the child. However, as the number and similarity of words that must be represented increases, there is greater pressure to develop a more systematic encoding of the relevant distinctions. Insofar as the same phonological representations subserve both comprehension and production (as in the current framework), the emergence of more segmental representations through pressures on comprehension should also manifest in production.

Another class of time-course phenomena concerns the order in which skilled performance is achieved for various phonological units (e.g., vowels, consonants, and consonant clusters). A coarse-grained example is the finding that the proportion of consonants in intentional/imitative utterances is low early on, and increases as the number of words produced increases (Vihman et al., 1985; Bauer, 1988; Roug, Landberg, & Lundberg, 1989). Our framework does not account for this directly, but there are two factors that bias vowel-like articulations during early imitation and intentional naming in our model. First, our articulatory representations have a built-in bias towards the center value of each degree of freedom, mimicking a physical bias of least articulatory effort. This bias causes oral constriction to be somewhat open (i.e., vowel-like) by default. Second, when the model begins to compare its own acoustics with those of the adult, it learns quickly that it must vocalize in order to produce most types of sounds. This coarse learning precedes learning the more complicated articulatory relationships involved in producing consonants. The combination of these two factors creates an early bias towards vowel-like sounds during imitation and intentional speech, which is overridden as the system gains greater control of the articulatory degrees of freedom.

Another, finer-grained example is that labials are produced more often in the first word productions than in babbling (Boysson-Bardies & Vihman, 1991). In learning to produce labial consonants (compared with more orally internal articulations), infants can use the visual feedback of labial articulations in addition to feedback derived from acoustics (see Vihman, 1996). The acoustic level of representation in our model includes a visual dimension that corresponds to jaw openness, which is most active for labial articulations. The additional visual information for labials should cause words with primarily labial articulations to be produced accurately sooner than other words because the error signal from labial productions is richer than from less visible articulations.

A similar finding is that children master the production of stops before fricatives and affricates (Menn & Stoel-Gammon, 1995). In the physical mapping from articulation to acoustics in our model, the range of oral constriction values that produces frication is much smaller that the range that produces the component events of a plosive sound (i.e., closure followed by release). Second, mandibular oscillation during babbling produces a bias favoring closure-release sequences which approximate plosives. This, in turn, causes the forward model to learn the articulatory-acoustic relationship for plosives before fricatives. The forward model will thus provide more accurate articulatory feedback for plosives than for fricatives as the system learns to produce words.

Phonetic Content of Speech Errors

Detailed analyses of speech errors have provided some of the most important constraints on theories of phonological development. Perhaps the most basic and widespread types of errors that children make during the early stages of word production are ones of simplification or reduction. Menn and Stoel-Gammon (1995) listed three such error types that we address in this section: 1) stops are substituted for fricatives; 2) consonant clusters are reduced to single consonants; and 3) voiceless initial stop consonants are deaspirated

(e.g., TOE is heard as DOE).

That plosives are mastered before fricatives (see above) is relevant to the first point. If the production system cannot activate the articulatory representations precisely enough to produce fricatives, then plosives are the likely alternative. With respect to consonant reduction, both acoustic and articulatory factors may be involved. The acoustic similarity of BOND and BLOND, for example, is very high, which means that learning distinct phonological representations for them will be difficult. The similar representations that such contrasts will drive may be sufficiently distinct for driving different semantics, but not different articulations. On the articulatory side, vowels and consonants constrain somewhat complementary sets of articulatory degrees of freedom. Consequently, consonant clusters permit relatively less coarticulation—and hence entail greater difficulty—compared with transitions between consonants and vowels (also see Kent, 1992). Also, the bias to produce simple over complex onsets may, in part, be due to their generally higher frequency of occurrence in speech. Finally, the deaspiration of initial unvoiced stops may be explained by articulatory difficulty (i.e., a least-effort bias; see above).

Conclusion

In conclusion, we propose a distributed connectionist framework for phonological development in which phonology is a learned internal representation mediating both comprehension and production, and in which comprehension provides production with error feedback via a learned articulatory-acoustic forward model. An implementation of the framework, in the form of a discrete-time simple recurrent network, learned to comprehend, imitate, and intentionally name a corpus of 400 monosyllabic words. Moreover, the speech errors produced by the network showed similar tendencies as those of young children. Although only a first step, the results suggest that the approach may ultimately form the basis for a comprehensive account of phonological development.

References

Atal, B. S., Chang, J. J., Mathews, M. V., & Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique. *Journal of the Acoustical Society of America, 63*, 1535–1555.

Bauer, H. (1988). The ethological model of phonetic development: I. *Clinical Lingistics and Phonetics, 2*, 347–380.

Benedict, H. (1979). Early lexical development: Comprehension and production. *Journal of Child Language, 6*, 183–201.

Berko, J., & Brown, R. (1960). Psycholinguistic research methods. In P. H. Mussen (Ed.), *Handbook of research methods in child development* (pp. 517–557). New York: Wiley.

Bernhardt, B. H., & Stemberger, J. P. (1997). *Handbook of phonological development.* New York: Academic Press.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of the perceptual re-organization for speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 345–360.

Boysson-Bardies, B. de., & Vihman, M. M. (1991). Adapatation to language: Evidence from babbling of infants according to target language. *Journal of Child Language, 67*, 297–319.

Boysson-Bardies, B. de., Vihman, M. M., Roug-Hellichius, L., Durand, D., Landberg, I., & Arao, F. (1992). Material evidence of infant selection from target language: A cross-linguistic phonetic study. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications.* Timonium, MD: York Press.

Chauvin, Y. (1988). *Symbol acquisition in humans and neural (PDP) networks.* PhD thesis, University of California, San Diego.

Davis, B. L., & MacNeilage, P. F. (1995). The articulatory basis of babbling. *Journal of Speech and Hearing Research, 38*, 1199–1211.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283–321.

Dodd, B. (1975). Children's understanding of their own phonological forms. *Quarterly Journal of Experimental Psychology, 27*, 165–173.

Eberhard, K. M., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research, 24*, 409.

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

Ferguson, C. A., & Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language, 51*, 419–439.

Fry, D. B. (1966). The development of the phonological system in the normal and deaf child. In F. Smith, & G. A. Miller (Eds.), *The genesis of language*. Cambridge, MA: MIT Press.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence, 40*, 185–234.

Houde, J. F. (1997). *Sensorimotor adaptation in speech production.* PhD thesis, Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences.

Huttenlocher, J. (1974). The origins of language comprehension. In R. L. Solso (Ed.), *Theories in cognitive psychology*. Hillsdale, NJ: Erlbaum.

Ingram, D. (1976). *Phonological disability in children.* London: Edward Arnold.

Jordan, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the 8th Annual Conference of the Cognitive Science Society* (pp. 531–546). Hillsdale, NJ: Erlbaum.

Jordan, M. I. (1992). Constrained supervised learning. *Journal of Mathematical Psychology, 36*, 396–425.

Jordan, M. I. (1996). Computational aspects of motor control and motor learning. In H. Heuer, & S. Keele (Eds.), *Handbook of perception and action: Motor skills.* New York: Academic Press.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science, 16*, 307–354.

Jusczyk, P. W. (1986). Toward a model of the development of speech perception. In J. S. Perkell, & D. H. Klatt (Eds.), *Invariance and variability in speech processes.* Hillsdale, NJ: Erlbaum.

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

Jusczyk, P. W., Pisoni, D. B., & Mullennix, J. (1992). Some consequences of stimulus variability on on speech processing by 2-month old infant. *Cognition, 43*, 253–291.

Kent, R. D. (1992). The biology of phonological development. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications.* Timonium, MD: York Press.

Kent, R. D., Mitchell, P. R., & Sancier, M. (1991). Evidence and role of rhythmic organization in early vocal development in human infants. In J. Fagard, & P. H. Wolff (Eds.), *The development of timing control and temporal organization in coordinated action.* Oxford: Elsevier Science.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development, 70*, 340–349.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Ladefoged, P. (1993). *A course in phonetics.* Orlando, FL: Harcourt Brace.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Liberman, A. M. (1996). *Speech: A special code.* Cambridge, MA: MIT Press.

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

Lieberman, P., & Blumstein, S. E. (1988). *Speech physiology, speech perception, and acoustic phonetics.* Cambridge, England: Cambridge University Press.

Lindblom, B. (1992). Phonological units as adaptive emergents of lexical development. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications.* Timonium, MD: York Press.

Lindblom, B., MacNeilage, P. F., & Studdert-Kennedy, M. (1984). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, & Ö. Dahl (Eds.), *Explanations for language universals.* Berlin: Mouton.

Lively, S. E., Pisoni, D. B., & Goldinger, S. D. (1994). Spoken word recognition: Research and theory. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 265–301). New York: Academic Press.

Locke, J. L. (1995). Development of the capacity for spoken language. In P. Fletcher, & B. MacWhinney (Eds.), *The handbook of child language* (pp. 278–302). Oxford: Blackwell.

MacNeilage, P. F. (in press). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences.*

MacNeilage, P. F., & Davis, B. L. (1990). Acquisition of speech production: The achievement of segmental independence. In W. J. Hardcastle, & A. Marchal (Eds.), *Speech production and speech modelling.* Dordrecht: Kluwer Academic.

Markey, K. L. (1994). *The sensorimotor foundations of phonology: A computational model of early childhood articulatory and phonetic development* (Technical Report CU-CS-752-94). Boulder: University of Colorado, Department of Computer Science.

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

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375–407.

McClelland, J. L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes, 4*, 287–335.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198*, 75–78.

Menn, L. (1978). Phonological units in beginning speech. In A. Bell, & J. B. Hooper (Eds.), *Syllables and segments.* Amsterdam: North-Holland.

Menn, L. (1983). Development of articulatory, phonetic, and phonological capabilities. In B. Butterworth (Ed.), *Language production*, Vol. 2 (pp. 3–50). New York: Academic Press.

Menn, L., & Stoel-Gammon, C. (1995). Phonological development. In P. Fletcher, & B. MacWhinney (Eds.), *The handbook of child language* (pp. 335–359). Oxford: Blackwell.

Menyuk, P., & Menn, L. (1979). Early strategies for the perception and production of words and sounds. In P. Fletcher, & M. Garman (Eds.), *Language acquisition: Studies in first language development.* Cambridge, UK: Cambridge University Press.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review*, *76*, 165–178.

Moskowitz, B. A. I. (1973). The acquisition of phonology and syntax: A preliminary study. In K. J. J. Hintikka, J. M. E. Moravcsik, & P. Suppes (Eds.), *Approaches to natural language*. Dordrecht: Reidel.

Nittrouer, S. (1993). The emergence of mature gestureal patterns is not uniform: Evidence from an acoustic study. *Journal of Speech and Hearing Research*, *30*, 959–1172.

Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research*, *32*, 120–132.

Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G. Yenicomshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology 1: Production.* New York: Academic Press.

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., & Jordan, M. I. (1995). Goal-based speech motor control: A theoretical framework and some preliminary data. *Journal of Phonetics, 23*, 23–35.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 37–42). Hillsdale, NJ: Erlbaum.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of naming and lexical decision. *Language and Cognitive Processes, 12*, 767–808.

Plaut, D. C., & Kello, C. T. (in preparation). *A distributed connectionist approach to phonological development.* Manuscript in preparation.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56–115.

Reznick, J. S., & Goldfield, B. A. (1992). Rapid change in lexical development in comprehension and production. *Developmental Psychology, 28*, 406–413.

Roug, L., Landberg, I., & Lundberg, L. J. (1989). Phonetic development in early infancy: A study of four Swedish children during the first eighteen months of life. *Journal of Child Language, 16*, 19–40.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Sedivy, J. C., Tanenhaus, M. K., & Spivey-Knowlton, M. J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632.

Smith, B. L. (1973). *The acquisition of phonology: A case study.* Cambridge, U.K.: Cambridge University Press.

Snyder, L. S., Bates, E., & Bretherton, I. (1981). Content and context in early lexical development. *Journal of Child Language, 6*, 565–582.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*, 217–257.

Stoel-Gammon, C., & Cooper, J. A. (1984). Patterns of early lexical and phonological development. *Journal of Child Language, 11*, 247–271.

Studdert-Kennedy, M. (1987). The phoneme as a perceptomotor structure. In A. Allport, D. MacKay, W. Prinz, & E. Scheere (Eds.), *Language perception and production.* New York: Academic Press.

Studdert-Kennedy, M. (1993). Discovering phonetic function. *Journal of Phonetics, 21*, 147–155.

Thelen, E. (1981). Rhythmical behavior in infancy: An ethological perspective. *Developmental Psychology, 17*, 237–257.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action.* Cambridge, MA: MIT Press.

Treiman, R., & Zukowski, A. (1996). Children's sensitivity to syllables, onsets, rimes, and phonemes. *Journal of Experimental Child Psychology, 61*, 193–215.

Van Orden, G. C., & Goldinger, S. D. (1994). Interdependence of form and function in cognitive systems explains perception of printed words. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1269.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review, 97*, 488–522.

Vihman, M. M. (1993). Variable paths to early word production. *Journal of Phonetics, 21*, 61–82.

Vihman, M. M. (1996). *Phonological development: The origins of language in the child.* Oxford: Blackwell.

Vihman, M. M., Maken, M. A., Miller, R., Simmons, H., & Miller, J. (1985). From babbling to speech: A re-assessment of the continuity issue. *Language, 61*, 397–445.

Vihman, M. M., & Miller, R. (1988). Words and babble at the threshold of lexical acquisition. In M. D. Smith, & J. L. Locke (Eds.), *The emergent lexicon: The child's development of a linguistic vocabulary.* New York: Academic Press.

Vihman, M. M., & Velleman, S. L. (1989). Phonological reorganization: A case study. *Language and Speech, 32*, 149–170.

Wagner, K. R. (1985). How much do children say in a day? *Journal of Child Language, 12*, 475–487.

Walley, A. C. (1993). The role of vocabulary development in children's spoken word recognition and segmentation ability. *Developmental Review, 13*, 286–350.

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.

Werner, H., & Kaplan, D. (1984). *Symbol formation: An organismic-developmental approach to language and the expression of thought.* Hillsdale, NJ: Erlbaum.

Williams, R. J., & Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation, 2*, 490–501.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). Forward dynamic models in human motor control: Psychophysical evidence. In G. Tesauro, D. S. Tourtezky, & T. K. Leen (Eds.), *Advances in neural information processing systems 7* (pp. 43–50). Cambridge, MA: MIT Press.