A SENSORSIMOTOR MODEL OF EARLY CHILDHOOD PHONOLOGICAL DEVELOPMENT

Kevin L. Markey

CU-CS-695-93

University of Colorado at Boulder
DEPARTMENT OF COMPUTER SCIENCE

A SENSORIMOTOR MODEL OF EARLY CHILDHOOD PHONOLOGICAL DEVELOPMENT

CU-CS-695-93 December 1993

Kevin L. Markey

Department of Computer Science
University of Colorado at Boulder
Campus Box 430
Boulder, Colorado 80309-0430 USA

Abstract

By the time they are two-years old, children typically have a productive vocabulary of at least 200 words, only one year after having spoken their first word. Despite this impressive achievement, children's speech only roughly approximates adult speech. Intriguingly, their errors show a systematic pattern which we may not dismiss as merely a matter of poor performance. Their error patterns demonstrate an evolving competence, a developing phonological system.

Longitudinal linguistic studies which have characterized many properties of early childhood phonological development observe development as a sequence of changes in linguistic structure. We wish to explain the structure of linguistic change, integrating a learning component capable of accounting for the evolution of linguistic knowledge. Connectionist methods are the obvious candidate, but most models in this area focus on morphological aspects of phonology which occur late in development. A relatively complete sensorimotor model (Laboissiere 1992) learns basic articulatory skills of speech. But it does not demonstrate the foundation of phonology: the ability to recombine basic sounds.

We face two dilemmas. Phonological competence may not be acquired without articulatory skill, and articulatory skills may only be acquired in the context of larger, more abstract phonological patterns. But should we choose a hierarchical system of cognitive and motor control, we face the problem of how to control a system which operates at multiple time scales when stimuli which seem to drive the system are continuously changing.

Our solution is to enlist the perceptual system to segment and categorize auditory feedback such that discrete perceptual events regulate the timing of the more abstract level of phonological control. We distribute articulatory skill among many experts, each specialized to produce a basic sound. The role of phonological control is two-fold: (1) to ensure that the distribution of sounds made by articulatory specialists matches the distribution of sounds in the linguistic environment, and (2) to activate them in a sequence to compose a more elaborate sound.

We propose solutions to several other computational dilemmas, and we propose a methodology to test the model's ability to demonstrate the phenomena of early childhood phonological development. Results are summarized for portions of the model which have been implemented.

For abstracts or reprints of more recent reports on this research, contact the author via email:
markey@cs.colorado.edu.

A Sensorimotor Model of Early Childhood Phonological Development
A Proposal

Kevin L. Markey
Department of Computer Science and Institute of Cognitive Science
University of Colorado, Boulder

1. Motivation

Within their first six months of life, children start to recognize prosodic patterns and vowels unique to their native language (Jusczyk 1992, Kuhl et al. 1992), and their babbling starts to show characteristics of adult speech (Oller & Lynch 1992). A mere two months later, they recognize consonant-vowel syllables unique to their linguistic environment (Werker et al. 1981, Werker 1992). By the end of their first year, children can recognize about fifty words (Benedict, 1979), their babbling now approximates many of the sounds in their native tongue (Vihman et al. 1986, de Boysson-Bardies et al. 1989), and they are speaking their first real words (Benedict, 1979). In the second year, their productive vocabulary mushrooms to several hundred words. Despite this impressive achievement, children will not have accurately mastered all sounds. Their speech is filled with errors of commission and omission, only roughly approximating adult speech (see Menn 1978, Ingram 1989). For example, many children transform adult fricatives like [f] and [s] into stops like [b] and [t]; others will substitute approximants [w] and [j] for liquids [l] and [r]. Sounds will often be dropped from consonant clusters like [spr] and [skw]. They might assimilate a sound in one part of a word with a similar sound in another part such that "dog" becomes [gOg]. Even more intriguing, such errors show a systematic pattern which we may not dismiss as simply a matter of poor performance (Smith 1973, Menn 1971). Instead, the error patterns demonstrate an underlying competence which slowly evolves.

Longitudinal linguistic studies reveal the changing compositional structure of children's speech, children's systematic errors relative to adult speech and how their errors change over time (e.g., Smith 1973, Menn 1971), phonetic trends evident in babble (e.g., Vihman et al. 1986), and unruly behavior which resists a clean analysis (Menn & Matthei 1992). In addition, comparative anatomy identifies constraints on the developmental course of certain motor skills (Kent & Murray 1982) and psycholinguistic experiments reveal children's emerging perceptual categorization of speech sounds (e.g., Jusczyk 1992, Werker & Pegg 1992, Kuhl et al. 1992, Grieser & Kuhl 1989).

Traditionally (Saussure 1916/1966), these longitudinal accounts would be viewed as diachronic — a description of how linguistic structure changes historically or developmentally (McNeill 1987) — and not synchronic — a description of linguistic structure at an instant in time. But at its heart, this traditional view of change as a series of instantaneous snapshots is synchronic. Rather than viewing language acquisition as merely a sequence of changes in linguistic structure, we wish to characterize the structure of linguistic change. Such an account of phonological development must integrate a learning component and must provide a principled explanation of how phonological competence evolves. Connectionist computation is an obvious candidate for this task.

Several connectionist models which learn phonological phenomena have been proposed. Most are limited in scope, such as those intended to explain children's acquisition of the English past tense (Rumelhart & McClelland 1986, MacWhinney & Leinbach 1991, Daugherty & Seidenberg 1992) or article declension in German (MacWhinney et al. 1989). Some are not intended as developmental accounts; they use connectionist methods to discover internal representations which explain strictly adult data (e.g., Touretzky & Wheeler 1990, Hare 1990). These models typically start with a perceptual representation which corresponds to the abstract distinctive features attributed to adult phonological competence (Shillcock et al. 1992). This assumption may be justified to the extent that the phenomena studied are an abstract elaboration of a largely pre-existing cognitive subsystem. But it is too strong an assumption if one's goal is to explain the development of that subsystem.

Our goal is to build and explore a developmental model of the phonetic and articulatory foundations of phonology which explains the phenomena of early childhood speech perception and production. We model a relatively complete sensorimotor system, intending to capture the interaction of perception and production as the system evolves.

A relatively complete sensorimotor and connectionist model of vocal tract control has been built (Laboissiere et al. 1990, Laboissiere 1992). It learns jaw, tongue, and lip motions necessary to pronounce a sequence of vowels. It makes no assumptions about phonetic features. The teaching signal is a trajectory of formant frequencies corresponding to the target vowel sequence. Errors in formant frequencies are inverted into articulatory errors by a forward model trained during a period of synthetic babbling (Jordan & Rumelhart 1992). Once trained, the model simulates skilled human behavior in classic bite-block and coarticulatory experiments. Were the model intended as a story of phonological development — which it is not — its principal weakness would be its inability to compose previously learned sound segments in new combinations, precisely the shortcoming of any model of articulatory competence which limits itself to motor skill.

We are caught on the horns of a dilemma. Phonological competence may not be acquired or demonstrated without articulatory skill, and articulatory skills are acquired in the context of larger, more abstract phonological patterns. Articulatory skill and phonological competence engage in a contrapuntal dance. Without acknowledging both, we cannot account for the basic facts of phonological development, neither the relative articulatory difficulty of some sounds nor the recombination of sounds in unique utterances.

We view speech as a two-level problem of motor and cognitive control, as a hierarchy of sequential decision tasks. But, once we admit to multiple levels of cognitive or motor control, we buy ourselves another dilemma — how it is possible to control a system which operates at multiple time scales when internal and external stimuli which seem to drive the system are continuously changing. Our solution is to enlist the perceptual system to segment and categorize auditory feedback such that discrete perceptual events regulate the timing of the more abstract level of phonological control. Furthermore, we enforce a fairly strict separation between the two control levels. This is not due to a fundamental commitment to modularity but rather because of the computational requirements of the problem.

We thus propose a sensorimotor model of early childhood phonological development. Its principal contribution is a hierarchical model of control in which the upper, more abstract level of control depends on the segmentation of time and acoustic feedback by a categorical perceptual system and on a heuristic to ensure the correct credit assignment between control levels. We also introduce an exploration strategy to pace the complexity of sounds being learned and thus increase the likelihood that the two control levels converge to a solution. Unlike Laboissiere's model, learning is not supervised by a teacher; rather, it is guided by internally generated reinforcement signals (see also Montague et al. 1993). We assume that all behavior is goal-directed, and unlike other theories, we make no qualitative distinction between babble and speech.

There is no phonetic-feature system of representation built into our model's auditory perception. Instead, auditory perception must discover the important categories of speech sounds in the linguistic environment by a simple competitive learning process. This process is assisted by a linguistically relevant segmentation algorithm, but segmentation criteria are strictly acoustic.

Control takes advantage of the degrees of freedom and constraints built into the vocal tract's anatomy. Articulatory motions are organized as gestures which resemble the dynamics of a critically damped spring system (Browman & Goldstein 1989). But articulatory gestures are chosen by an articulatory controller with a parallel architecture (Markey 1994). Without this innovation, learning articulatory control would be computationally impractical.

In the next section, we describe the proposed model in detail. Section 3 establishes the experimental methodology, including the selection of data for training, what behaviors we hope to observe, how the model explains them, and how we will measure them. Section 4 briefly explains shortcomings in the model and opportunities for future research to correct them. Section 5 summarizes the results we have thus far obtained in implementing about sixty percent of the proposed model and presents my plan to complete the work and dissertation.

2. The Model

2.1. Overall Architecture

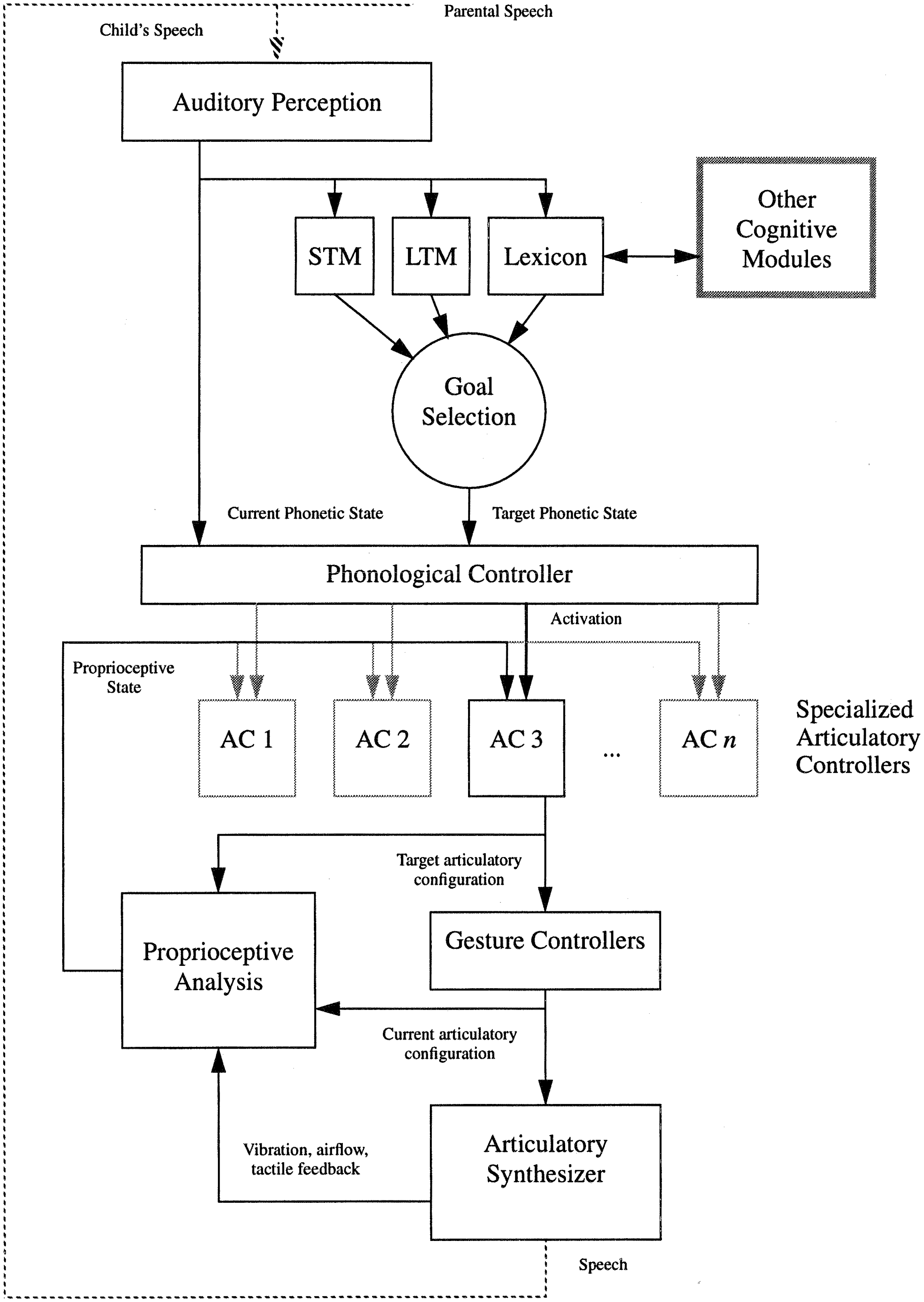
The model's input (see figure 1) is a synthetic acoustic signal representing the parent's speech or the child's own vocalizations. This acoustic input is segmented and converted into a categorical phonetic representation by the *auditory perception* module. A cumulative record of a complete utterance is maintained in such a way as to uniquely represent any one-syllable (or simple two-syllable) morpheme. Phonetic representations are stored in *short-term memory* or the *lexicon* from which they may be retrieved by *goal selection* to become a target utterance. The phonetic representation of the child's acoustic feedback also serves to report the model's progress during the course of an utterance and to determine the accuracy of the final result.

Control is hierarchical and is separated into three domains — phonetic, articulatory, and gestural. The *phonological controller* attempts to follow a phonetic trajectory which achieves the phonetic goal. It does so by activating a sequence of articulatory controllers. It does not attempt to control the articulatory configuration directly; it has no direct knowledge of the articulatory or proprioceptive state; it selects articulatory controllers which can be expected (in some statistical sense) to achieve the phonetic state it desires; it rewards the articulatory controllers which do its bidding.

An *articulatory controller* attempts to traverse an articulatory and proprioceptive trajectory which optimally meets the needs of its phonological client. It does so by initiating a sequence of articulatory gestures. It has no direct knowledge of the phonetic state and no direct knowledge of certain gestural details. It only observes the articulatory, tactile, and proprioceptive state of the articulatory system and reinforcement signals from its client. There are several articulatory controllers. Each is a specialist capable of generating all or part of a recognizable sound. Each articulatory controller has a parallel architecture with one subagent per articulatory degree of freedom.

Gestures chosen by the articulatory controller are executed by the *gesture controllers.* There is one gesture controller per articulatory degree of freedom. Each is a fixed procedure which moves an articulator from its current state to a target equilibrium state in a trajectory resembling that of a

Figure 1: Control Architecture



critically damped spring of a given stiffness. The outputs of all gesture controllers together define the vocal tract's configuration. The *articulatory synthesizer* — an adaptation of Haskins Laboratories' ASY (Rubin et al. 1981) — converts the vocal tract configuration into a synthetic acoustic signal, which is then processed by the model's auditory perception to complete the feedback loop. Proprioceptive state, input for the articulatory controller, is a compilation of outputs from the active articulatory controller, gesture controllers, and the articulatory synthesizer which are analyzed by the *proprioceptive analysis* module.

Plastic components of the model are phonetic categorization processes, phonological and articulatory controllers. Reinforcement learning shapes production of speech. Categories of sound patterns recognized by auditory perception are discovered by a simple competitive learning process.

2.2. Critical Properties of the Model

Event-driven timing of phonological control is the key to phonological compositionality. If control decisions must be made continuously or at each tick of a discrete clock, there is no way to achieve true compositionality. Introducing segmented categorical auditory perception is the prerequisite. However, some level of control cannot help but occur in continuous or finely-grained discrete time. The articulatory gymnastics necessary to achieve recognizable acoustic events require exquisite timing. This is the role of the lower-level articulatory controllers, which learn the requisite fine-grained temporal timing of articulatory gestures.

This is an efficient division of labor. A longer utterance, while within the realm of possibility, is more difficult to learn or plan as a low-level motor skill simply because of its greater combinatorial complexity. By breaking up a long utterance into smaller segments, each of which may be learned individually or in other contexts, we simplify the learning task. The articulatory controller design allows several to be chained together to complete an articulatory puzzle under certain circumstances, even if they were initially trained to accomplish more specific tasks in other contexts.

Computationally, the job of the articulatory controller is to approximate elemental sound segments in the model's environment. The phonological controller's job is to approximate longer utterances by activating a sequence of articulatory controllers, once the latter have mastered some simple sounds. Before then, however, the phonological controller's job is to build a model which estimates the probability with which each articulatory controller generates some sound in the environment and which matches the statistical distribution of sounds in the linguistic environment. Thus does the phonological controller enforce a diversity of articulatory capabilities and learns which articulatory controllers are best suited to generate various target sounds.

This scheme — combining low-level skill acquisition with skill composition — is not guaranteed to converge. Indeed, too complex a target may lead to a deadlock between articulatory search and phonological search such that neither converges. An exploration strategy is introduced to avoid such deadlocks. Phonological complexity is paced by the gradual increase in the complexity of sounds recognized by auditory perception. Systematic goal selection, sound preference, and other strategies more efficiently stage articulatory exploration in ways resembling children's sound exploration strategies. All are subject to experimental manipulation.

There is another critical computational issue. Articulatory control has so many degrees of freedom that a traditional central control architecture is computationally impractical. Thus, we introduce a parallel control architecture in which each articulatory subcontroller is dedicated to each degree of freedom in articulatory control space, but where all subcontrollers share the same proprioceptive state data and global reinforcement signal. The subcontrollers cooperatively solve the articulatory control problem.

Grounding the model is a one-to-one correspondence between the model's fundamental articulatory and perceptual building blocks — articulatory gestures and phonetic segments. This is based on the observation that static phonetic segments correspond to the beginnings or ends of gestures and that peak spectral transitions correspond approximately to the zenith of articulatory gestures.

2.2.1. Explaining basic phenomena

A more concrete example will help explain some of the features and properties of the model. We consider the common substitution of stop consonants for fricatives. Similar stories can be told for other phenomena (see section 3.3., p. 25).

The relative difficulty of pronouncing various sounds is related to their relative combinatorial complexity and the proprioceptive cues available to guide articulatory control. For example, stop consonants require less complex motion, less precise control, and fewer proprioceptive cues than fricatives. Fricatives will simply take longer to learn than stops, and until they are mastered, the likelihood of accurately rendering a fricative will be considerably lower than that of a phonetically similar stop consonant. Despite the possibly greater accuracy achieved by choosing an articulatory controller specialized in fricative sounds, the phonological controller will instead activate the articulatory controller specialized in stop consonant motions because it is more likely to generate a nearby sound, and thus it predicts a higher reward.

2.3. Auditory Perception

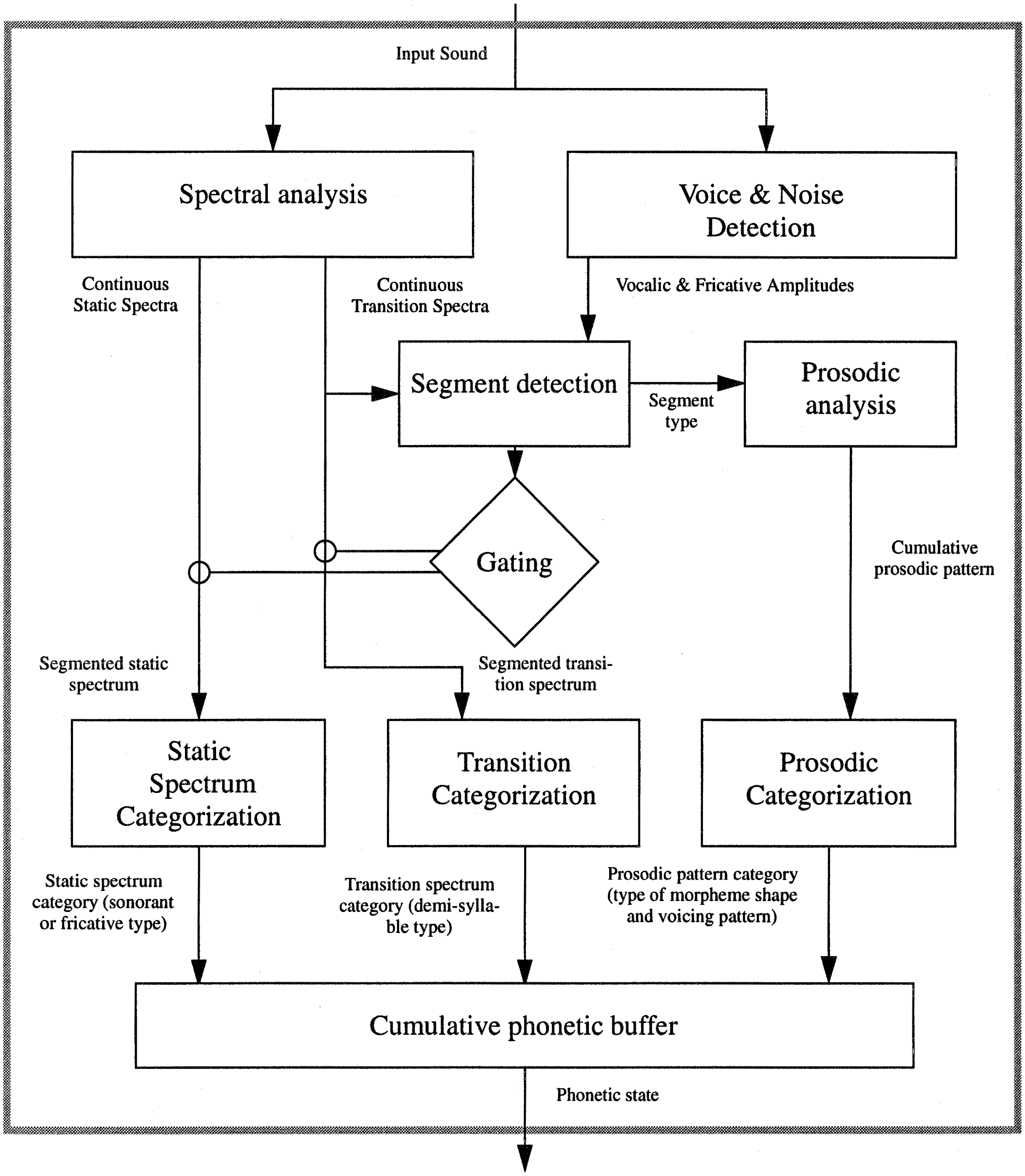
Auditory perception processes incoming acoustic data at four levels in three domains (see figure 2). The first level extracts static and transition spectra, plus voice and noise amplitudes. The second level segments the signal by first evaluating the magnitude of spectral transitions and inspecting periods of voice, noise, and silence. It then samples spectral, transition, and prosodic features at each appropriate segment. The third level clusters these features into statistically relevant categories, activating prototype representations closest to the stimulus. Cumulative prosodic categories correspond to morpheme shape and voicing pattern detected thus far during the utterance. Static spectrum categories correspond to vowel, sonorant, or fricative types. Transition spectrum categories correspond to demi-syllable types. The fourth level accumulates a trace of prototype activations over the entire course of the utterance, representing its complete phonetic trajectory. The module's output, the phonetic state, remains unchanged until the next segment is detected. The representation should uniquely represent any one-syllable utterance except those which diverge too far from the norm.

2.3.1. Spectral Analysis

Spectral analysis starts with a 256-frequency power spectrum sampled once every 8 msec and performs the following analyses. An illustration of the resulting analysis appears in figure 3.

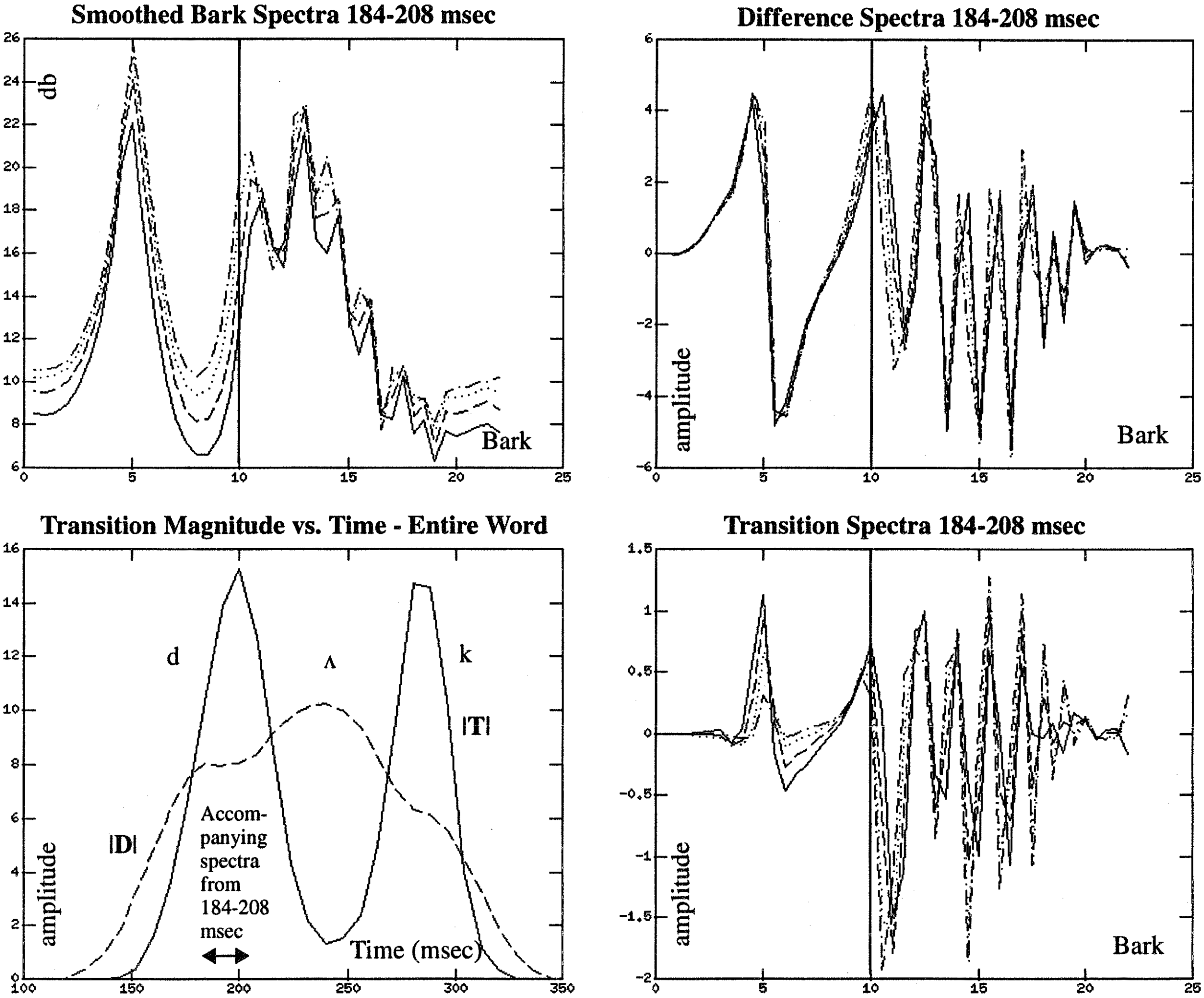
- **Static Bark spectrum:** The input power spectrum is sampled evenly over the frequency domain. To better approximate the frequency response of the human ear and to reduce the size of the parameter space, we convert frequencies to bark (Zwicker 1961) and divide the input spectrum among 44 coarsely coded 1.0-bark wide detectors, b_i for $1 \leq i \leq 44$, evenly spaced one per 0.5 bark between 0.5 and 22.0 bark.

Figure 2: Auditory Perception Module


Input Sound
Spectral analysis
Voice & Noise Detection
Continuous Static Spectra
Continuous Transition Spectra
Vocalic & Fricative Amplitudes
Segment detection
Segment type
Prosodic analysis
Gating
Cumulative prosodic pattern
Segmented static spectrum
Segmented transition spectrum
Static Spectrum Categorization
Transition Categorization
Prosodic Categorization
Static spectrum category (sonorant or fricative type)
Transition spectrum category (demi-syllable type)
Prosodic pattern category (type of morpheme shape and voicing pattern)
Cumulative phonetic buffer
Phonetic state

- **Smoothed static spectrum:** To filter short-term noise, we apply a Gaussian filter with a standard deviation of 2 frames (16 msec) in the temporal dimension (Bradshaw & Bell 1991).
- **Static difference spectrum:** To measure detector amplitudes on a relative, not absolute, scale, we generate a "difference" spectrum **D**, as the difference in amplitude between neighboring bark detectors, $D_i = b_i - b_{i-1}$ for $i > 1$.
- **Transition spectrum:** To approximate the first time derivative of each difference detector's amplitude, we compute the difference between detector amplitudes at times *t* and *t*-1. The resulting transition spectrum **T** is simply $T_i(t) = D_i(t) - D_i(t-1)$.

Figure 3: Spectra and Onset Transition for "duck"



Smoothed Bark Spectra 184-208 msec

Difference Spectra 184-208 msec

Transition Magnitude vs. Time - Entire Word

Transition Spectra 184-208 msec

Transition magnitude vs. time (lower left) shows L_1 norms of transition spectra **T** and difference spectra **D** for the entire audible portion of "duck." Other plots each show 4 spectra sampled at 184, 192, 200 and 208 msec. The 10-bark detector is highlighted. The peak near 10 bark is the 2nd formant. Its frequency is decreasing. This is captured in the transition spectra (lower right) by a positive value in the direction of the formant's motion (to the left) and a negative value behind its motion (to the right).

2.3.2. Segment Detection

The model segments the acoustic input during periods of voicing and frication by observing changes in the transition magnitude, an adjusted L_1 norm of the transition spectrum. It samples static spectra when the transition magnitude is minimal or transition spectra when transition magnitude is maximal. When the model cannot detect the spectrum of the acoustic input (during silence or when speech sounds are barely audible), it identifies each continuous period of silence, noise, or barely audible voicing as a segment. For each detected segment, the submodule transmits the type of segment to the prosodic analysis submodule. It transmits either the sampled static or sampled transition spectrum, depending on the segment type. (See figure 3 for a graph of transition magnitude over a CVC syllable plus samples of transition spectra near the peak transition.)

- **Transition magnitude:** The L_1 norm of transition spectrum minus the L_1 norm of the difference spectrum is the transition magnitude. The former captures temporal changes in both total amplitude and spectral shape, the latter captures only changes in amplitude. To detect each extreme transition we wish to detect only when change in spectral shape is at an extremum. The result is averaged over three 8-msec frames.
- **Extreme transition detection:** Each local maximum in transition magnitude above a fixed threshold and each local minimum below a fixed threshold is identified. Local extrema less than 35 msec apart are presumed to be part of the same segment whose location we assume to lie halfway between.
- **Average static and transition spectrum:** We average the transition spectra over the 50 msec period centered at the point of extreme transition.

The representation, although linguistically motivated, is based on strictly acoustic criteria. The result is not a traditional phonetic or phonological segment like a "phoneme." To be sure, the stored prototypes are based on the statistical properties of parental speech, but they are not "distinctive" in the phonemic sense. The approach is suggested by the observation that spectral transitions provide the most salient portion of the speech signal for consonants or syllable recognition (Furui 1986). Maximum transitions correspond to demi-syllables or diphthongs. Minimum transitions correspond to relatively steady-state portions of vowels, nasals, approximants, and fricatives. Barely audible noise is aspiration or a stop release. Barely audible voicing is the voice-bar accompanying a voiced oral stop consonant. And silence represents either the closure of a voiceless stop consonant or morpheme boundaries.

2.3.3. Prosodic analysis

The prosodic analysis submodule performs a very simple analysis of morpheme shape and voicing pattern by concatenating segment types identified by the segment detection submodule. The model does not detect pitch or stress. We use the term "prosodic" in the sense of suprasegmental phonetic patterns (e.g., Waterson 1971) rather than the more usual sense of phrase prosody.

2.3.4. Categorization

A set of "prototypes" is stored in long term memory representing those transition spectra, static spectra, and prosodic patterns recognized as relatively distinct according to an appropriate distance measure and statistically important according to their relative frequency in parental and child speech. For each token speech segment in each perceptual domain, we measure its distance from existing prototypes. Each prototype is activated by an amount inversely related to its distance from the token segment. If the token's distance from every prototype is too great, we may add a new prototype. To measure the distance between token and prototype segment for static or transition spectra, we use the inverse correlation distance (Pomerleau 1993). To compare the token prosodic string with prosodic prototypes we use dynamic time warping and a cosine distance metric for each segment type vector. This three-part categorization avoids the conflation of spectral and prosodic differences.

2.3.4.1. Prototype acquisition and loss

If a token is too distant from existing prototypes, a new prototype is added probabilistically, using the current token as the prototype. Such probability is proportional to the token's distance to its nearest neighbor, but is low enough to discourage a proliferation of gratuitous categories. A prototype is lost with a probability (1) inversely related to its frequency of detection, (2) inversely related to its distance from neighboring prototypes, and possibly (3) inversely related to its relevance for phonological control. Because phonetic state is encoded as prototype activation, phonological control cannot be reliable until prototype competition stabilizes.

2.3.4.2. Stipulated and emergent properties of staged categorization

We expect that phonetic categories will be learned in order of increasing complexity, a function of their frequency of occurrence and mutual phonetic distances. This is one factor which paces the complexity of sounds that the model may attempt to reproduce, helping to bootstrap motor learning by phonological and articulatory controllers. However, in experimental manipulations of the model, we plan to introduce additional constraints, including (1) explicit manipulations of parental sounds, and (2) constraints on the number or type of phonetic categories which may be stored as prototypes at some given stage of development (see sections 2.8.2 and 3.3.3).

2.3.5. Phonetic buffer and utterance representation

The activation of phonetic prototypes represents only the momentary phonetic state. The cumulative phonetic buffer maintains a trace vector of all prototype activations accumulated during the course of an utterance. It is this cumulative record which becomes the output of auditory perception — the "phonetic state". For example, the sound [pIg] would be represented by activation of the prototype onset transition for [bI] (the transitions for [pI] and [bI] are not distinguished), the coda transition for [Ig], the prototype for the steady-state spectrum corresponding to [I], and a prosodic string for silence, aspiration, onset transition, vocalic steady-state, coda transition, a barely audible voicebar, and the burst of noise representing the consonant release. "Big" [bIg] is distinguished from [pIg] only by a voicebar which replaces [pIg]'s initial silence and aspiration.

2.3.5.1. Uniqueness of representation for syllable and morpheme recognition

It is possible to uniquely represent any single syllable with this representation. The relative order of spectral features is implicit in transition spectra because they represent overlapping demi-syllables (as above). By adding static spectral prototypes to the representation, we capture nasal, fricative, and stand-alone vocalic data missing from a strictly dynamic spectral analysis. Their order relative to other features is usually encoded in spectral transitions. The prosodic representation can disambiguate most remaining questions of feature order, but is primarily necessary to distinguish voiced, voiceless, and fricative sounds. It is also possible to distinguish some polysyllabic utterances, especially reduplicated syllables like /mama/. However, the proposed representation is primarily aimed at a solution for single-syllable utterances and syllable reduplication.

2.3.5.2. A structured phonetic state which uniquely represents any morpheme.

A more elaborate phonetic state which resembles a tensor-product binding (Smolensky 1990) of prosodic (or syllable frame) roles and spectral feature fillers is able to uniquely represent any morpheme. Instead of a simple vector, we form a matrix by taking the cross-product of all possible spectral features and prosodic categories. After each phonetic segment is detected, we activate the unit corresponding to its spectral category and the prosodic pattern observed thus far in the utterance. Unfortunately, such a representation could become huge. It would also be cumbersome to modify as phonetic and prosodic categories are added or lost. At present, implementation of a structured phonetic state is beyond the scope of this project, not being crucial for testing its central hypotheses.

2.3.5.3. Characteristics of code

Though the recognition code is linguistically motivated, its phonetic features are learned, not assumed. Prototype matching yields a coarse coding scheme which retains some distributed properties but emphasizes segmental and categorical properties of speech perception. It is computationally simple and amenable to parallel implementation.

However, the combinatorial complexity of syllable and/or morpheme structure is large, requiring substantial memory resources for prototype storage and large computational costs for sequential implementation of prototype matching. Limiting ourselves to only 11 English vowel types and only 3 basic places-of-articulation, there are 66 demi-syllable types, not including diphthongs, vowel-glide, or glide-vowel types or additional variations caused by introducting nasality. Add to this a large number of syllable types. This concern may be ameliorated by artificially limiting the adult corpus of sounds or the model's capacity, forcing it to ignore less frequent sound patterns.

The code does not decompose phonetic segments into distinctive features, nor demi-syllables into consonant and vowel. Feature decomposition and other strategies may be necessary for a more compact, less complex representation, but are beyond the scope of this project.

I have already implemented the spectral analysis, segmentation detection, and spectral distance algorithms. Simulations were able to discriminate and classify synthetic CV syllables much like those we will use for the adult corpus (Markey & Bell 1993).

2.4. Memory

The model has three varieties of memory. Short-term memory stores only the most recent utterance until the start of a new utterance is detected. Long-term memory is more permanent and more selective. It is implemented as a table in which a new utterance is stored if not already present, but only to the extent that its phonetic components are recognized as well-formed by segment categorization; that is, an utterance is added with a probability roughly proportional to its prototype activations.

The lexicon is a form of long-term memory in which "meaning" is also stored. Here "meaning" is limited to a few semantic and pragmatic features useful for guiding goal selection (e.g., animacy, people, personal relationships, food, names, etc.). Lexicon entries are added in a way which simulates the natural acquisition of receptive vocabulary. We shall implement the lexicon as a simple elaboration of long-term memory. A more complete model might introduce a more complete semantic component or a learning component which adds lexical items to the extent that their phonetic representation reliably predicts some cognitive event. These are beyond the scope of this project. The lexicon is important only as a criterion in goal selection and as a small contribution to reinforcement in some circumstances.

2.5. Goal Selection

There are two basic goal strategies which the model can pursue.

- Imitate a speech sound in the environment (parent's or one's own most recent utterance).
- Reproduce the utterance type stored in long-term or lexical memory.

The goal selection module will copy a phonetic representation from the appropriate memory to the phonological controller. Goal and phonetic target probabilities will be a function of the model's developmental stage or some experimental manipulation. During the babbling stage, imitation of self or parents will be emphasized. As long-term and lexical memory expand, they will provide a greater proportion of targets. Preferences for various sounds or lexical items may be weighted by phonetic or semantic factors, according to experimental manipulation or systematic exploration strategy. Such preferences are a key component of motor learning bootstrapping strategies intended to overcome deadlocks between phonological and articulatory controller searches.

2.6. Phonological control

The phonological controller's task is to choose a sequence of articulatory controllers such that the actual phonetic state comes to match the target phonetic state. It has two inputs, the current phonetic state and the target phonetic state. Its only output is the activation of some articulatory control specialist, and it guides articulation only indirectly, by its choices of articulatory controller.

2.6.1. Q-learning

The phonological controller is a Q-learning agent (Watkins 1989). Q-learning is a reinforcement algorithm for discovering an extended plan of action which maximizes the cumulative net long-term reward received by an agent as result of its actions. In general, the Q-agent incrementally learns a function, $Q(x, a)$, which for every environmental state x evaluates the expected utility of performing each possible action a. Optimally, the Q-agent chooses the most highly valued action. As it learns the Q-function, however, the agent experiments with possibly suboptimal actions. As an optimal plan is learned, $Q(x, a)$ (or the Q-value, as it is sometimes called) comes to equal the expected value of the cumulative net reward which would be gained by performing action a in state x and by following the optimal plan of action in subsequent steps.

Q-learning has been shown to converge to an optimal plan for a finite Markov sequential decision task under a number of specific conditions (Watkins & Dayan 1992). Our implementation is adapted from Lin (1992), using a multi-layer network to estimate expected utility for each state and action. Despite evidence that convergence is not guaranteed when Q-values are estimated by such function approximators (Watkins 1989, Thrun & Schwartz 1994), the practice is widespread in order to take advantage of their generalization properties, especially for large state spaces.

2.6.2. Phonological controller action selection and error correction

The neural network implementing the phonological controller has sufficient inputs to accommodate the current phonetic state **x**, target phonetic state **p**, and one output unit for each articulatory controller. Hidden units are sigmoidal; output units are linear. The target phonetic state **p** is chosen by the goal selection module at the start of each trial and remains constant throughout the trial. The controller considers and chooses its next action once at the start of each new trial and afterwards only when a change in the phonetic state **x** is detected, which is a function of auditory perception. Its i-th output is the estimated utility $Q(\mathbf{p}, \mathbf{x}, a_i)$ of choosing articulatory controller a_i given the current and target phonetic states. Action a is chosen and articulatory controller a is activated according to a Boltzmann distribution $p(a_i | \mathbf{x}, \mathbf{p})$ across all possible actions A. Temperature T determines the randomness of the action selection and is varied during learning according to some annealing schedule, the chosen goal strategy, and a bootstrapping strategy.

$$p(a_i | \mathbf{x}, \mathbf{p}) = e^{Q(\mathbf{p}, \mathbf{x}, a_i)/T} \Big/ \sum_{a_k \in A}^{n} e^{Q(\mathbf{p}, \mathbf{x}, a_k)/T} \tag{1}$$

If the phonological controller is optimally trained, this may be interpreted as the probability that articulator a_i will generate the next phonetic segment in a sequence which culminates with the current and target phonetic states equal. If constrained such that the phonological controller may choose one and only one articulator during the course of a trial, then this may be interpreted as the probability that articulator a_i will generate a sound represented by the target phonetic state.

The Q-function is next evaluated only when auditory perception next detects a phonetic segment and the phonological controller observes a new phonetic state **y**. At this time, the controller will also observe the scalar cost C incurred by the previous action and scalar reinforcement R. Before the next action is chosen, we compute the temporal difference error E (Sutton 1988) in our prediction of action a's utility and use it to update the controller's parameters. Our new predicted utility is the sum of cost C, reinforcement R, and the maximum utility among all actions in our new state **y**. The error is the difference of this sum and $Q(\mathbf{p}, \mathbf{x}, a)$, the utility of the chosen action.

$$E = R + C + \max_{a_k \in A} \{Q(\mathbf{p}, \mathbf{y}, a_k)\} - Q(\mathbf{p}, \mathbf{x}, a) \tag{2}$$

The network is adjusted by back propagating the square of the error for that output unit corresponding to action a. No error is back propagated through other output units. The calculation uses activations stored at the time the action was taken.

2.6.3. Reinforcement of the phonological controller

During the course of an utterance, we compare the current phonetic state with the target phonetic state. A trial ends when the goal state is satisfied, after some time-out period determined by a simple respiratory model, if some impossible articulatory configuration is encountered, or (optionally) if a mismatch occurs, that is, if the utterance is recognized as one of the same class of sounds as the target but not of the same type. Various reinforcement schedules are possible and require some experimentation. They range from a positive reinforcement for achieving the goal and negative reinforcement for mismatches or time-outs, to a graded signal which peaks for an exact match and falls away exponentially with the token's distance from the target. The phonological controller receives a reward only to the extent that the whole utterance matches the target utterance.

2.7. Articulatory control

There are several articulatory controllers. Each learns to become a specialist in producing all or part of some recognizable sound. To produce a sound, the active articulatory controller governs the motions of jaw, tongue, lips, velum, lungs, and glottis to adjust the shape of the vocal tract and generate air turbulence or vibration. To produce some desired sound pattern, the controller observes and traverses a proprioceptive trajectory without any direct knowledge of sound. At this level of control, the length of time necessary to observe and analyze auditory feedback is too great to be of any use. Its only knowledge of sound is via the reinforcement signals it receives for the sounds it produces.

2.7.1. Parallel architecture of articulatory controller and its relationship to gesture controllers

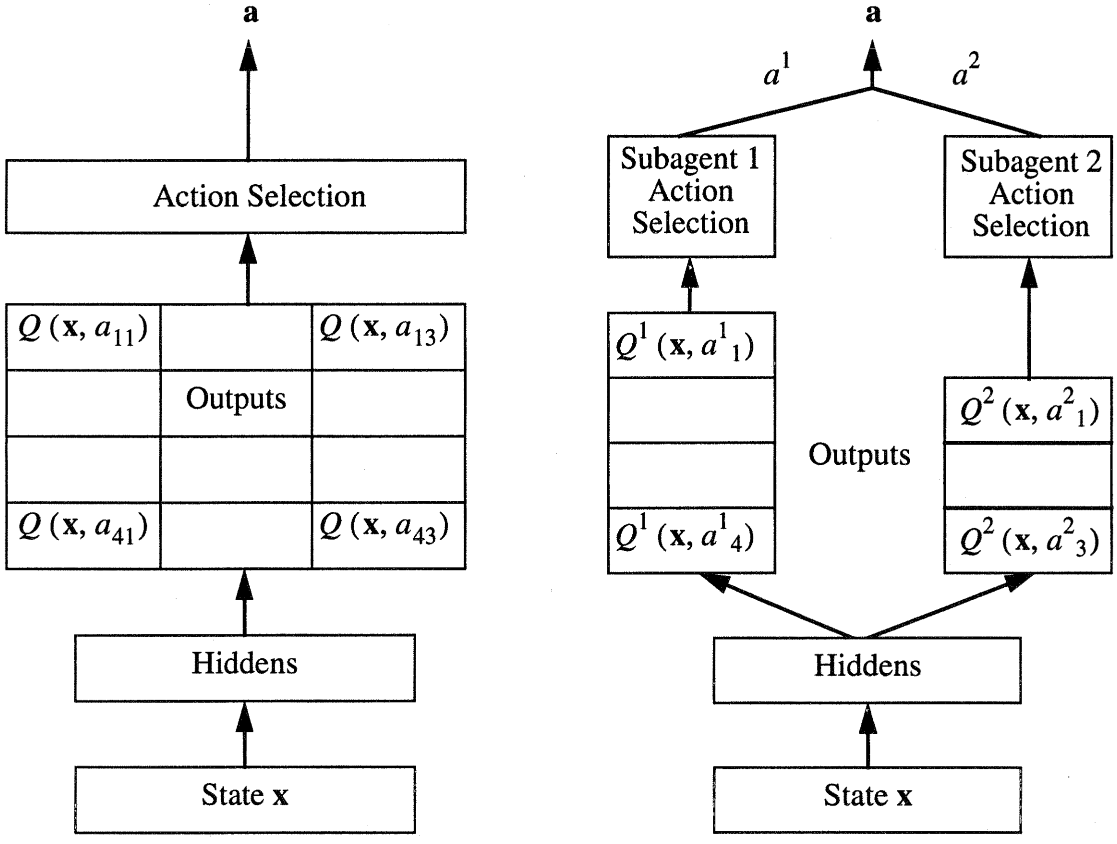
Motions of the six articulators occur along 12 degrees of freedom (see 2.10) organized as a set of articulatory gestures. Each gesture is the motion of one articulator in one dimension from its current position to a chosen equilibrium position along a trajectory conforming to the motion of a critically damped spring (Browman & Goldstein 1989; see 2.9). The articulatory controller does not choose each point through which a vocal tract articulator will pass during a gesture, only the equilibrium position. The gesture controller realizes the actual motion (see sections 2.6.2, 2.9).

Each articulatory controller resembles a parallel computer. Rather than a single agent controlling all twelve articulatory dimensions, twelve agents each control a single dimension. Each subagent is assigned permanently to a single dimension (e.g., jaw angle, tongue tip displacement, lip protrusion) and its corresponding gesture controller. And each subagent chooses among a range of gestural target values spaced evenly (or roughly so) along the entire range of motion of its articulatory dimension plus a "null" action which has no effect on the current trajectory.

The articulatory controller employs the parallel quasi-independent Q-agent architecture introduced by Markey (1994). We do so because the standard method would be impractical, requiring at least 50 million outputs to specify all possible vocal tract configurations with sufficient resolution (one for each combination of gesture targets across all 12 degrees of freedom; see figure 4).

Figure 4: Standard and Parallel Q-Agent Architectures Compared

We portray the explicit representation of Q-values for all possible action combinations in a standard Q-agent on the left (for a small 4x3 action space). On the right we portray the sparse representation of Q-values in the quasi-independent architecture for the same action space. Subagents make action choices by comparing Q-values within only a single articulatory dimension. The controller's action is composed from each component choice.



Standard Q-Agent Architecture

Quasi-independent Parallel Q-Agent Architecture

In the parallel-Q architecture, one subagent is dedicated to each articulatory dimension. The collective behavior of all subagents replaces the behavior of a single agent. All subagents share the same input — the proprioceptive state — which describes the recent and current state of the vocal tract. The algorithm is presented in greater detail below.

2.7.2. Time granularity of articulatory control

At each time step (corresponding to 8 msec of real world time) and for each degree-of-freedom, the gesture controller updates the current gesture's trajectory, and the corresponding subagent chooses the next gesture. A gesture which matches the previously chosen gesture has no effect. If a new gesture is chosen during a refractory period immediately following the start of the previous gesture, it replaces the old gesture with a probability quadratically proportional to the progress of the old gesture.

2.7.3. Articulatory controller action selection and error correction

Details of the parallel, quasi-independent Q-agent architecture used for articulatory controllers appear in Markey (1994). We summarize the algorithm here.

The action space **G** from which the articulatory controller chooses action (gesture target) **g** has n degrees of freedom, each defining a separate subspace, G^j, j=1..n, where $\mathbf{G} = G^1 \times \ldots \times G^n$. The actions in each subspace represent the range of articulatory gestures in one articulatory dimension. The method for each subagent is similar to that used by a single Q-agent (Watkins 1989); however, instead of learning a complete Q-function, each subagent explores only those actions in its own subspace and learns only the corresponding portion of the Q-function. There is no explicit representation of the entire Q-function; together subagents must cooperatively explore the entire action space.

- There is one network whose input is proprioceptive state **u**. It has $\sum |G^j|$ outputs, organized into n groups, each group representing a subagent, one for each degree of freedom j. Subagent group j has only $|G^j|$ outputs. At time t, each computes the estimated value $Q^j(\mathbf{u}, g^j)$ of some action $g^j \in G^j$.
- Each subagent selects an action g^j in its subspace according to the Boltzmann distribution. The action seen by the environment is the n-tuple $\mathbf{g} = (g^1, \ldots, g^n)$ constructed of the actions chosen independently by each network. Action selection by each subagent is independent of selections by other subagents. The Boltzmann distribution is computed only over actions in G^j and only over outputs in group j.

$$p(g_i^j | \mathbf{u}) = e^{Q(\mathbf{u}, g_i^j)/T} \Big/ \sum_{g_k^j \in G^j}^{n} e^{Q(\mathbf{u}, g_k^j)/T} \qquad (3)$$

- Execution of action **g** results in new state **v**, and all subagents observe the same cost c, and reinforcement r. The temporal difference error E^j is computed *for each subagent group j.*

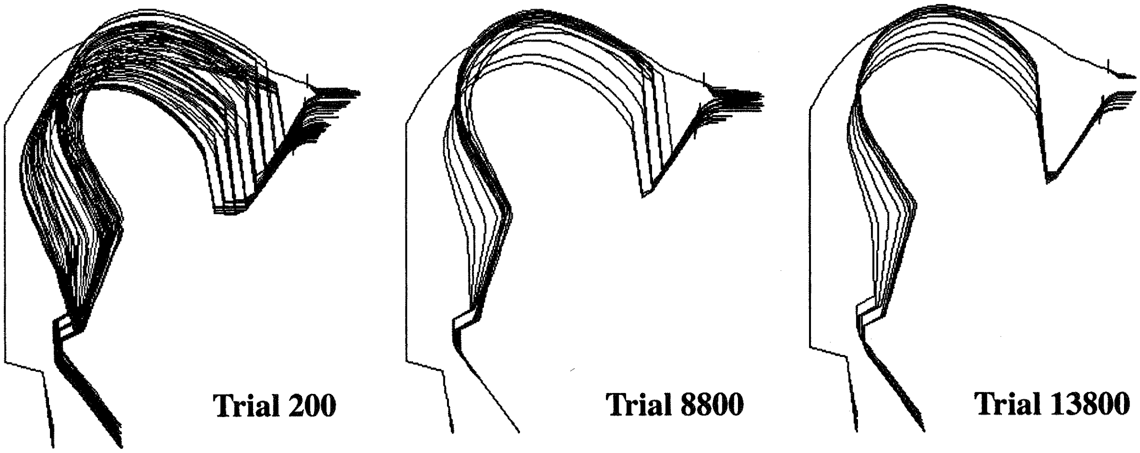
$$E^j = r + c + \max_{g^j_k \in G^j} \{Q^j(\mathbf{v}, g^j_k)\} - Q^j(\mathbf{u}, g^j) \qquad (4)$$

- *For each subagent group j* the square of this error is back propagated through the output unit corresponding to the action g^j selected by group j at time t.

Experiments with this architecture have demonstrated its ability to learn primitive CV syllables in as little as 2 to 8 hours of experimentation (in the simulated time frame of the child), starting with a completely naive controller.

Figure 5: Articulatory controller learning velar contact and release

Contact of velum and release by tongue body learned within 13,800 trials.



2.7.4. Reinforcement of articulatory controller

An articulatory controller receives rewards and penalties only for those phonetic events detected while it is active. Reinforcement is accumulated over this period but does not accrue until control is transferred to a different articulatory controller or until the trial times out; cost is accrued at each time step. A reward is received if the distance between the current phonetic state and the target phonetic state is narrowed. No reward results if the distance increases. A penalty results if a phonetic segment is generated which is not activated in the target phonetic state representation. For example, if our goal is to pronounce the word "big" and the onset demi-syllable [bI] has already been detected, the active articulatory controller receives a reward if the demi-syllable [Ig], receives no reward if a second [bI] demi-syllable token is detected, and receives a penalty if the demi-syllable [ga] is detected.

2.8. Relationships between phonological and articulatory controllers

In this section we discuss in greater detail the computational roles of phonological and articulatory controllers. Though the specialization of the latter yields certain efficiencies, joint training on complex utterances need not converge without some exploration strategy to bias their joint search of articulatory space. We evaluate such exploration strategies. We also compare the model's control architecture with related models. Finally, we explain under what circumstances articulatory controllers may be chained together by the phonological controller. The relationship of the articulatory and phonological controllers outlined thus far is summarized in table 1.

2.8.1. Computational roles and relationships between control levels

Phonological and articulatory controllers face several computational tasks. First, there is the problem of approximating elemental sound patterns which recur in the environment. This is the role of

Table 1: Articulatory and phonological controllers compared

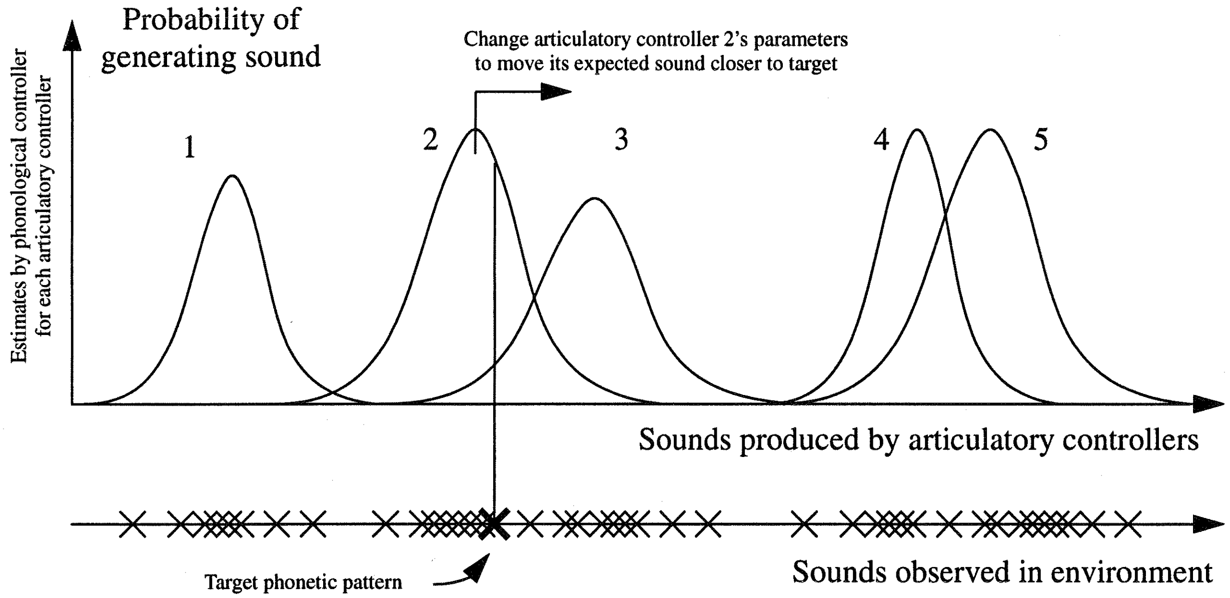
Dimension	Articulatory controller	Phonological controller
Quantity	Many	1
State	Proprioceptive state	Current & target phonetic states
Source of state information	Vocal tract configuration, tactile and proprioceptive feedback, proprioceptive events detected and stored by proprioceptive analysis	Segmented, categorical auditory perception of current utterance and stored representation of target utterance.
Action	Parallel gesture choice and activation	Articulatory controller choice and activation
Granularity	New state computed, action chosen every 8 msec; but action choice is event-driven.	Event-driven. When each new phonetic segment is detected and phonetic state changes.
Learning and control algorithm	Parallel Q-learning	Q-learning
Architecture	Quasi-independent Q agents	Single parameterized Q agent
Computational analogy	Server	Client

the articulatory controllers. Second, the problem of recombining these simple sounds to approximate more complex, possibly unique sound patterns falls to the phonological controller. Once the articulatory controllers have mastered some simple sounds of appropriate diversity, recombining them is a relatively simple task. Until then, each has a much more difficult task.

It is crucial to ensure an appropriate diversity of sounds, to ensure that the distribution of sounds generated by the articulatory controllers comes to match the distribution of sounds which occur in the environment. This is the phonological controller's task. It builds a model which estimates the probability with which each articulatory controller generates some sound in the environment. Following this model, it probabilistically activates one of articulatory controllers closest to each target sound. The chosen articulatory controller is shaped by reinforcement learning to better approximate the target sound. Finally, the phonological controller updates its own model based on the performance of the chosen articulatory specialist. The model that the phonological controller uses is simply its Q-function, which together with the Boltzmann distribution (1) defines the relative likelihood that each articulatory controller will generate the sound pattern represented by the target phonetic state. Its Q-function is adjusted by reinforcement learning as the phonological controller attempts various matches.

This is illustrated (figure 6) with a comparison to the Gaussian mixture model. Given some distribution of speech sounds (scattergram in figure), we assume that a collection of articulatory controllers generates them (mass density curves in figure), not unlike assuming that some mixture of Gaussians generates a set of arbitrary observations (Nowlan 1991). Given an observation, the Gaussian mixture model would ask "which Gaussian generated it?" That is, it would calculate the conditional probabilities that each Gaussian generated the observation and then incrementally update the means and variances of the Gaussians to more closely model the data. Given some

Figure 6: Building a model of articulatory controllers to match linguistic environment


Probability of generating sound
Estimates by phonological controller for each articulatory controller
Change articulatory controller 2's parameters to move its expected sound closer to target
1
2
3
4
5
Sounds produced by articulatory controllers
Target phonetic pattern
Sounds observed in environment

sound sample from the distribution (e.g., the bold "X"), the phonological controller asks which articulatory controller is most likely to have generated it. It calculates the conditional probabilities of the articulatory controllers per its Q-function and the Boltzmann distribution. It activates an articulatory controller per this distribution (e.g., controller 2 in figure). Though we know in advance how to directly adjust means and variances of Gaussians, we can only indirectly update the parameters of each articulatory controller by reinforcement learning. When we do so, the chosen articulatory controller's "mean" sound stochastically and incremently moves closer to the target sound. The phonological controller adjusts its model of articulatory controller behavior by adjusting its own Q-function as a function of how well the chosen articulatory controller generated the target sound, adjusting curve 2 (adjustment not shown in figure).

2.8.2. An exploration strategy to bootstrap the architecture

One can imagine the circumstances under which this learning scheme will fail to converge. The model may prematurely attempt a complex phonetic target without having first mastered the component sounds. Human children encounter the same problems. Before mastering its component sounds, Jacob experiments with an untold number of variations in "thankyou" for six months before finally giving up, never once accurately rendering the complex utterance (Menn data 1976). The model could also pursue an unselective goal strategy, only exacerbating the already undifferentiated articulatory controllers when, early in development, Boltzmann temperatures are high and distributions are broad and flat.

One solution is an exploration or "bootstrapping" strategy which stages the complexity of utterances attempted by the child or otherwise limits the complexity of search. We propose two approaches. One is a simple experimental manipulation by which we break the model's training into two phases. First we limit training to the simplest demi-syllables. Only then do we expose the model to more complex utterances. The second bootstrapping strategy is more complete, recruiting several existing facilities already proposed for the model, but adding several new constraints.

Dayan, P. and Hinton, G.E. (1993). Feudal reinforcement learning. In Hanson, S.J. et al. (Eds.), *Neural Information Processing Systems 5:* 271-278.

de Boysson-Bardies, B., Halle. P., Sagart,L. and Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. *Journal of Child Language 16:* 1-17.

de Boysson-Bardies, B. and Vihman, M.M. (1991). Adaptation to language: evidence from babbling and first words in four languages. *Language 67:* 297-319.

de Boysson-Bardies, B., Vihman, M.M., Roug-Hellichius, L., Durand, C., Landberg, I., and Arao, F. (1992). Material evidence of infant selection from the target language: a cross-linguistic phonetic study. In C.A. Ferguson, L. Menn, and C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications,* pp. 369-391. Parkton, MD: York Press.

Durand, J. (1990). *Generative and Non-Linear Phonology.* London: Longman.

Evarts, E.V. (1971). Feedback and corollary discharge: a merging of the concepts. *Neurosciences Research Program Bulletin 9*(1): 86-112.

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

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

Gathercole, S.E. and Baddeley, A.D. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language 29:* 336-360.

Goldstein, U.G. (1980). *An articulatory model for the vocal tracts of growing children.* Ph.D. Thesis. Cambridge, MA: MIT, Dept of Electrical Engineering and Computer Science.

Grieser, D. & Kuhl, P.K. (1989). Categorization of speech by infants: support for speech-sound prototypes. *Developmental Psychology 25:* 577-588.

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

Hare, M. (1990). The role of similarity in Hungarian vowel harmony: a connectionist account. *Connection Science 2:* 123-150.

Higgens, J.R. and Angel, R.W. (1970). Correction of tracking errors without sensory feedback. *Journal of Experimental Psychology 84:* 412-416. Monitor and correct own behavior in less than proprioceptive reaction time.

Huggins, A.W.F. (1964). Distortion of the temporal pattern of speech: interruption and alternation. *Journal of the Acoustical Society of America 36:* 1055-1064.

Ingram, D. (1974). Phonological rules in young children. *Journal of Child Language 1:* 49-64.

Ingram, D. (1989). *First Language Acquisition: Method, Description, and Explanation.* Cambridge University Press, Cambridge.

Jacobs, R.A., Jordan, M.I, Nowlan, S.J. and Hinton, G.E. (1991). Adaptive mixtures of local experts. *Neural Computation 3:* 79-87.

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

Jusczyk, P.W. (1992). Developing phonological categories from the speech signal. In C.A. Ferguson, L. Menn, and C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications,* p. 17-64. Parkton, MD: York Press.

Kelso, J.A.S., Tuller, B., Vatikiotis-Bateson, E. and Fowler, C.A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance* 10: 812-832.

Kent, R.D. and Bauer, H.R. (1985) Vocalizations of one-year-olds. *Journal of Child Language 12:* 491-526.

Kent, R.D. and Murray, A.D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *Journal of the Acoustical Society of America 72:* 353-365

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

Laboissiere, R. (1992). *Préliminaires pour une robotique de la communication parlée: inversion et contrôle d'un modèle articulatoire du conduit vocal.* Ph.D. Thesis. Grenoble: l'Institut National Polytechnique de Grenoble.

Laboissiere, R., Schwartz, J., and Bailly, G. (1990) Motor control for speech skills: a connectionist approach, In Touretzky, David S. et. al. (Eds.), *Connectionist Models: Proceedings of the 1990 Summer School,* pp. 319-327. San Mateo, CA: Morgan Kaufmann Publishers.

Lee, B.S. (1950). Effects of delayed speech feedback. *Journal of the Acoustical Society of America 22:* 824-826.

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning 8:* 293-321.

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

Macken, M.A. (1979). Developmental reorganization of phonology: a hierarchy of basic units of acquisition. *Lingua 79:* 11-49.

MacWhinney, B. (1991). *The CHILDES Project.* Hillsdale, NJ: Erlbaum.

MacWhinney, B., Leinbach, J., Taraban, R. and McDonald, J. (1989). Language learning: cues or rules? *Journal of Memory and Language 28:* 255-277.

MacWhinney, B. and Leinbach, J. 1991. Implementations are not conceptualizations: Revising the verb learning model. *Cognition 40:* 121-157.

Markey, K. and Bell, A. (1993). The peak transition spectrum as a compact, robust, and linguistically relevant speech code. Manuscript.

Markey, K. (1994). Efficient learning of multiple degree-of-freedom control problems with quasi-independent Q-agents. In Mozer, M.C. et al. (Eds.), *Proceedings of 1993 Connectionist Models Summer School,* pp. 272-279. Hillsdale, NJ: Erlbaum.

McNeill, D. (1987). *Psycholinguistics: A New Approach.* New York: Harper and Row.

Menn, L. (1971). Phonotactic rules in beginning speech. *Lingua 26:* 225-251.

Menn, L. (1976). *Pattern, control, and contrast in beginning speech: a case study in the development of word form and word function.* Ph.D. Thesis. University of Illinois. Published by Indiana University Linguistics Club.

Menn, L. (1978). Phonological units in beginning speech. In A. Bell and J.B. Hooper (Eds.), *Syllables and Segments,* pp. 157-171. North-Holland Publishing Co.

Menn, L. (1983). Development of articulatory, phonetic, and phonological capabilities. In B. Butterworth (Ed.), *Language Production,* vol. 2, pp. 3-50.

Menn, L. and Matthei, E. (1992). The "Two Lexicon" Account of Child Phonology: Looking Back, Looking Ahead. In C.A. Ferguson, L. Menn, and C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications,* p. 211-247. Parkton, MD: York Press.

Menn, L., Markey, K., Mozer, M. and Lewis, C. (1992). Connectionist modeling and the microstructure of phonological development: A progress report. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage and J. Morton (Eds.), *Developmental neurocognition: Speech and face processing in the first year of life.* The Netherlands: Kluwer Academic Publishers B.V.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America 53* (4): 1070-1082.

Montague, P.R., Dayan, P., Nowlan, S.J., Pouget, A. and Sejnowski, T.J. (1993). Using aperiodic reinforcement for directed self-organization during development. In *Neural Information Processing Systems 5:* 969-976.

Nowlan, S.J. (1990). Competing experts: an experimental investigation of associative mixture models. Technical Report CRG-TR-90-5. Toronto: University of Toronto, Department of Computer Science.

Nowlan, S.J. (1991). Maximum likelihood competitive learning. In D.S. Touretsky (Ed.), *Advances in Neural Information Processing Systems 2:* 574-582. San Mateo, CA: Morgan Kaufmann Publishers.

Oller, D.K. and Lynch M.P. (1992) Infant vocalizations and innovations in infraphonology: toward a broader theory of development and disorders. In C.A. Ferguson, L. Menn, and C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications,* pp. 509-536. Parkton, MD: York Press.

Oller, D.K., Wieman, L.A., Doyle, W.J., Ross, C. (1976). Infant babbling and speech. *Journal of Child Language 3:* 1-11.

Pomerleau, D.A. (1993). Input reconstruction reliability estimation. In Hanson, S.J. et al. (Eds.), *Neural Information Processing Systems 5:* 279-286.

Rubin, P., Baer, T. and Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of Acoustical Society of America 70* (2): 321-328.

Rumelhart, D.E. and McClelland, J.L. (1986). On learning the past tense of English verbs. In J.L. McClelland and D.E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the Microstructure of Cognition,* vol. 2, pp. 216-271. Cambridge, MA: MIT.

Saussure, F. de (1916/1966). *Course in General Linguistics.* New York: McGraw Hill.

Shaiman, S. (1989). Kinematic and electromyographic responses to perturbation of the jaw. *Journal of the Acoustical Society of America 86:* 78-88.

Shillcock, R., Lindsey, G., Levy J. and Chater, N. (1992). A phonologically motivated input representation for the modeling of auditory word perception in continuous speech. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society,* pp. 408-413. Hillsdale, NJ: Erlbaum.

Singh, S.P. (1992). Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning 8:* 323-340.

Smith, N.V. (1973). *The acquisition of phonology: a case study.* Cambridge: Cambridge University Press.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence 46:* 159-216.

Sutton, R.S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning 3:* 9-44.

Taylor, F.V. and Birmingham, H.P. (1948). Studies of tracking behavior II. The acceleration pattern of quick manual corrective responses. *Journal of Experimental Psychology 38:* 783-795.

Thrun, S. and Schwartz, A. (1994). Issues in using function approximation for reinforcement learning. In Mozer, M.C. et al. (Eds.), *Proceedings of 1993 Connectionist Models Summer School,* pp. 255-263. Hillsdale, NJ: Erlbaum.

Touretzky, D.S. and Wheeler, D.W. (1991). A computational basis for phonology. In D.S. Touretsky (Ed.), *Advances in Neural Information Processing Systems 2:* 372-379. San Mateo, CA: Morgan Kaufmann Publishers.

Vihman, M. (1978). Consonant harmony: its scope and function in child language. In J. Greenberg (Ed.), *Universals of Human Language,* v. 2, pp. 281-334. Stanford: Stanford University Press.

Vihman, M.M. and Velleman, S.L. 1989. Phonological reorganization: a case study. *Language and Speech 32:* 149-170.

Vihman, M.M. (1992). Early syllables and the construction of phonology. In C.A. Ferguson, L. Menn, and C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications,* p. 393-422. Parkton, MD: York Press.

Vihman, M.M., Ferguson, C.A. and Elbert, M. (1986) Phonological development from babbling to speech: common tendencies and individual differences. *Applied Psycholinguistics 7:* 3-40.

Vihman, M.M., Macken, M.A., Miller, R., Simmons, H. and Miller, J. (1985). From babbling to speech: a reassessment of the continuity issue. *Language 61:* 395-443.

Waterson, N. (1971). Child phonology: a prosodic view. *Journal of Linguistics 7:* 179-211.

Watkins, C.J.C.H. and Dayan, P. (1992). Technical note: Q-Learning. *Machine Learning 8:* 279-292.

Watkins, C.J.C.H. (1989). *Learning from delayed rewards.* Ph.D. Thesis. Cambridge, England: Cambridge University.

Werker, J.F., and Pegg, J.E. (1992). Infant speech perception and phonological acquisition. In C.A. Ferguson, L. Menn, and C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications,* p. 285-311. Parkton, MD: York Press.

Werker, J.F., J.H.V. Gilbert, K. Humphrey, and R.C. Tees. (1981). Developmental aspects of cross-language speech perception. *Child Development 52:* 349-355.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands. *Journal of Acoustical Society of America 33* (2): 248.

- Auditory perception learns to categorize simpler patterns first. This should already be an emergent property of phonetic categorization, but we may experimentally control how fast new categories are added or what types of categories are added.
- We adjust temperature annealing to constrain search. We set initial Boltzmann temperatures such that phonetic distributions of articulatory controllers show some modest differentiation yet are still plastic. Then, we "freeze" articulatory controllers once they become reliable, allowing others to experiment freely. Reliability may be measured as some threshold Q-value. Alternatively, we anneal articulatory controllers with a speed proportional to their Q-values when activated.
- Strategies and preferences for goal selection may take bootstrapping needs into account. Certainly, we may introduce some lexical preferences as a function of semantic and pragmatic features (section 2.4). We may experimentally introduce intrinsic sound preferences. More generally, we base goal selection in part on relative Q-values as estimated by the phonological controller at the start of each competing utterance. (Such estimates may be recorded as part of the lexicon or long-term memory.)

Bootstrapping strategies are expected to be idiosyncratic, a function of model history, stipulated preferences, and experimental manipulations.

2.8.3. Related models of hierarchical control

Our architecture bears some resemblance to several models of hierarchical learning or control. Feudal reinforcement learning (Dayan & Hinton 1993) is a true modular control hierarchy in which controllers are responsible for actions by subcontrollers and labor is delegated according to a non-adaptive, a priori territorial division. Composite Q-learning (Singh 1992) is an adaptation of the competing experts model (Nowlan 1990) which decomposes a complex sequential decision task into its elemental subtasks. Owing to the constraints of the domain, our architecture falls somewhere between feudal and CQ reinforcement learning in the space of possible models of hierarchical control.

A key distinction which drives other choices is whether to view the overall problem to be solved as a single sequential decision task or as a hierarchy of separate sequential decision tasks. Even if viewed as a single task, state information for a complex task must be organized at several levels. In choosing how to structure the task, one must ask whether the states of different levels differ in temporal (or spatial) granularity and/or in domain. For example, in feudal learning the states of each level describe the same domain (maze location), but differ in their territorial granularity. In our model, states of each level differ in domain (phonetic vs. proprioceptive state) and granularity. By granularity, we mean the time scale of inputs and actions, not the time scale of landmark events which occur sporadically in the course of continuous observations and behavior. The CQ-L algorithm observes its progress in accomplishing subgoals after each time step; its gating module must choose an expert Q module each time step. In contrast, our phonological controller chooses a new articulatory controller only with it observes a new phonetic state.

Division of labor may be territorial, temporal, or more generally, according to substates partitioned by some criterion. It may be adaptive or non-adaptive. Criterial state may alternatively be a categorical observation made in another domain, a learned criterion, or an a priori criterion. The phonological controller depends on its observation of the phonetic domain to divide articulatory responsibilities. The feudal maze criteria are non-adaptive, a priori territorial divisions.

One next chooses a method of distributing credit among experts or levels of control. The method depends in part on the structure chosen. Dayan and Hinton enforce reward hiding and inform each task manager what its conditions of satisfaction are. Singh credits temporal difference error according the posterior probability that an expert is chosen given the current state's value. Finally, one may apply various exploration strategies to constrain or guide search among levels of the hierarchy, or one may rely on the emergent properties of credit assignment to guide the search.

Table 2: Models of Hierarchical Reinforcement Learning Compared

Phonological/Articulatory Model	**Feudal Reinforcement Learning with Maze Task**	**Composite Q-Learning with Maze Task**
Interpreted as hierarchy of sequential decision tasks.	Interpreted as hierarchy of sequential decision tasks.	Interpreted as single SDT to be decomposed.
States at two levels differ in granularity and domain.	States at various levels differ in granularity only.	States at various levels differ in domain only.
Temporal division of subtasks by recognition of criterial phonetic state.	A-priori territorial division of state.	Temporal division of subtasks by recognition of subtask completion.
Adaptive division and assignment of articulatory controller.	Non-adaptive division and assignment of submanager or worker.	Adaptive division and assignment of expert.
Articulatory controller is a specialist on one task.	No specialization of workers to tasks. Worker is generalist.	Expert is a specialist on one task. Bias added to guide composite tasks.
Each articulatory controller sees the same undivided state.	Each worker sees subset of state.	Each expert sees the same undivided state. Bias network sees only task bits.
Phonological controller must decide how long to delegate subtask.	Subtask delegated until complete. Complete when manager's state changes.	Duration of delegation by gating module depends on detected progress in composite task.
Phonological controller activates articulatory controller with highest expected value.	Manager assigns task to preassigned worker.	Gating agent chooses expert with maximum likelihood of generating target Q-value.
Reward credited to active articulatory controller based on movement toward phonetic goal state during its activation.	Reward credited to chosen worker based on worker's success relative to task's known conditions of satisfaction.	Q-value TD error distributed between bias module & experts based on maximum likelihood of predicting target Q-value.
Various heuristics for bootstrapping specialization of articulatory controller.	No bootstrapping necessary because of non-adaptive division of state space.	No explicit bootstrapping.

2.8.4. Marketplace heuristic for distribution of rewards

Our task, which adaptively assigns jobs to workers, demands a mechanism for distribution of rewards which more closely resembles a marketplace than a feudal fiefdom. A labor marketplace retains a largely hierarchical organization, but rewards labor which meets needs determined by management. We require a reward distribution heuristic which assures a balance between "supply" and "demand" for articulatory services. We wish to reward those articulatory actions which fulfill a client need; we do not wish to reward excess services; but, we do not wish to penalize articulatory actions which exceed the demands of the task only because the manager (the phonological controller) purchased excess services. The phonological controller is rewarded on the basis of its success in reproducing a whole utterance (section 2.6.3., p. 13). The articulatory controller is rewarded on the basis of phonetic progress toward the target utterance only while it is active (section 2.7.4., p. 15).

2.8.5. Seamless transfer of control among articulatory controllers

Phonological compositionality requires the seamless integration of control as responsibility for motor control is transferred from one articulatory controller to the next. Moreover, learning the pronunciation of complex morphemes could be more efficient if we may freely piece together parts of previously learned articulatory puzzles to solve a new puzzle. For example, it would be advantageous if we could generate "mad" [m&d] out of previously learned articulatory patterns for "bad" [b&d] and "man" [m&n].

It is a fact of the articulatory system and a consequence of our model's design that the proprioceptive state is similar for steady-state phonetic segments shared across different utterances. The proprioceptive state corresponding to the articulation of the vowel [&] held in common by "mad", "bad", and "man" enables the seamless transfer of control between articulatory controllers which have previously mastered "man" and "bad". Likewise, multi-syllabic utterances joined by a consonant share not just the phonetic segment, but the proprioceptive state corresponding to the consonant. More generally, it is only necessary that corresponding proprioceptive states are in the same neighborhood, as determined by the generalization properties of the articulatory controller.

These properties are key to building the control architecture we have proposed, and they impose certain constraints on the model's design. The articulatory control domain must conform to the Markov property — the probability of moving to state $\mathbf{v}$ from state $\mathbf{u}$ given some action $\mathbf{g}$ is strictly a function of the current state $\mathbf{u}$ and action $\mathbf{g}$, not of any previous states. This is ensured by the design of gesture controllers, by information hiding between phonological and articulatory levels, and by the representation of proprioceptive state. Each articulatory controller must receive the same proprioceptive state information, which they do. Furthermore, an articulatory controller may not generate a representation of proprioceptive history which is not shared with other controllers. History must be compiled externally and must be shared equally among all controllers. In our model, history is compiled externally by the proprioceptive analysis module, and it is shared alike by all articulatory controllers. Their feedforward architecture avoids any private representation of history. A recurrent architecture here would create misleading internal representations of the proprioceptive state.

2.9. Gesture controllers

There is one gesture controller per articulatory dimension. Each gesture controller's input is the output of the articulatory controller's corresponding subagent — a target equilibrium position of the corresponding articulatory synthesizer control parameter. Its purpose is to generate a smooth trajectory which conforms to the motion of a critically damped spring from the articulator's current position to the chosen equilibrium position (Browman & Goldstein 1989). The speed of this motion is specified by a spring constant such that the motion approximates a range of speeds of the human vocal tract. The outputs of all gesture controllers together define a sequence of static vocal tract configurations which are interpreted by the articulatory synthesizer.

Gesture trajectories are updated at each discrete articulatory controller time step (corresponding to 8 msec of real world time). Once initiated, a gesture with a particular equilibrium position continues to its asymptotic conclusion unless interrupted by the articulatory controller with a new target equilibrium position. During the course of each gesture, the gesture controller tracks progress along its trajectory with a measure which roughly corresponds to "phase" in state space. By this and other means, timing of articulatory control is intrinsic. There is no clock, other than the intrinsic dynamics of each vocal tract articulator as implemented by gesture controllers.

2.10. Articulatory synthesizer

The articulatory synthesizer is an adaptation of Haskins Laboratories' ASY synthesizer (Rubin et al. 1981, Mermelstein 1973). At each discrete articulatory controller time step, it converts the static vocal tract configuration into a synthetic acoustic signal. The stream of such acoustic signals acts as the model's auditory output and feedback.

The vocal tract configuration is defined by control parameters for six articulators plus voicing and frication which vary over 12 degrees of freedom: jaw (joint angle), tongue body (distance and angle relative to the jaw hinge), tongue tip (length and angle relative to a point on the tongue body), lips (protrusion and separation), hyoid (anterior and superior distance relative to a fixed point), velum (degree of nasal opening), voicing amplitude (decibels), and frication amplitude (decibels). Voicing and frication control are unrealistically primitive. We add a simple sound source model which better approximates the indirect control of voicing and frication by replacing voicing and frication parameters with control parameters for lung expansion (arbitrary units) and glottal approximation (angle). From these input parameters, ASY computes the coordinates of each articulator, a mid-sagittal outline of the vocal tract and a frontal outline of the lips, the vocal tract's area function (its cross-sectional area from larynx to lips), the vocal tract's transfer function, and the components of the model's synthetic acoustic output: the power spectrum, voice and frication amplitudes (all functions of source characteristics and transfer function). Synthesized sound and graphics of the vocal tract are available for experimental observation of the model's behavior. Based on articulator coordinates, mid-sagittal outline, and cross-sectional areas we also compute signals which simulate tactile sensation.

Even with our extensions, ASY has weaknesses which limit the phenomena it can model. Its frication model is extremely primitive; modeled fricative sounds do not match corresponding human aspiration or fricatives. Accurately producing liquids [l] and [r] requires a three-dimensional vocal tract model; but ASY is two-dimensional. Thus, parameters for the liquids are not fully integrated. Jaw and lip motion are highly constrained, making it difficult to simulate labiodental [v] and [f] sounds. Finally, the synthesizer can not be configured to simulate vocal tract anatomies of different lengths and maturities. Thus, this model does not address the difficult problem of vocal tract normalization.

2.11. Proprioceptive analysis

Proprioceptive analysis recognizes certain proprioceptive events, and compiles the state and history of vocal tract motions from a number of sources. The resulting cumulative analysis becomes the proprioceptive state, common input for each articulatory controller and each subagent therein. It compiles:

- Current vocal tract configuration: a real vector of articulator positions, the outputs of gestural controllers, each scaled between 0 and 1.
- Target vocal tract configuration: a real vector of target articulator equilibrium positions, the outputs of articulatory controllers, each scaled between 0 and 1.
- Gesture phase: A representation of the current progress of each articulator through its current gesture, a measure roughly corresponding to phase in state space. Phase has a cyclic 6-unit representation starting with "at rest" and proceeding through "launch", "post-launch", "zenith", a sequence of "relaxation" phases, ending once again with "at rest".
- Touch detectors: Detectors for glottal vibration, airflow, vocal tract constriction, and touch at various regions in the vocal tract roughly corresponding to places of articulation.
- Touch and release event traces: We also detect and record a decaying trace memory of locations in the vocal tract which correspond to touch then release of contact.

3. Simulation and Experimental Methodology

An experimental test of the proposed model shares many of the methodological difficulties which face a linguist who observes and tries to characterize the phonological behavior of a child. Testing the real or model child is not a simple matter of passive training on some training corpus followed by a test of the child's generalization ability on a test set. The model was built as a whole sensorimotor system. It was also built to approximate the human child's exploratory behavior. Children are active explorers and learners, generating their own data. Not even the natural linguistic environment is passive; it consists of adult utterances responsive to the needs, interests, and speech of the child. Two human children or two models exposed to the same linguistic environment will demonstrate differences if only because of their divergent native endowments and their divergent histories.

Each individual child's behavior is remarkably idiosyncratic. We expect the same of model instances, even without modeling the full richness of the child-parent interaction. The only common denominator among individual children or model instances are general types of learned or innate behavior, types of errors, and general properties common to all phonological behavior. Our aim in testing the model's explanatory power will be to use linguistic and other methodologies to examine the model's behavior and compare it with known human behavior. This process and an analytic examination of the model and its properties may also reveal new questions about human behavior.

Analysis of the model's behavior could be an entire research project, given the richness of analytic methodologies available. We will choose among the techniques and measurements presented here as time and resources permit.

3.1. Available observational and experimental methodologies

Several methodologies have been used to observe and analyze human phonological development.

- Basic tools are naturalistic observation and detailed phonetic transcription (e.g., Smith 1973, Menn 1976), usually as part of a longitudinal study of one or more children.
- Phonetic transcriptions are supplemented by various linguistic structural analyses. These include rewrite rules which encode the systematic relationship of parent and child forms (Menn 1971, Smith 1973), prosodic structures or canonical forms which describe systematic gestalt patterns of sounds (Waterson 1971, Menn 1976, Macken 1979), phone trees (Ferguson & Farwell 1975, Ingram 1979), and various approaches to characterize irregular phenomena (Menn 1976, Menn & Matthei 1992). Phonological progress is reported as a sequence of stages, each described by its own structural analysis. A finer grained analysis is sometimes used to evaluate short-term changes in the pronunciation of a word or a class of words.
- Statistical studies of longitudinal or cross-linguistic trends and individual differences are used primarily to characterize babble and very early speech (de Boysson-Bardies et al. 1989, de Boysson-Bardies and Vihman 1991, Vihman et al. 1986).
- Experimental manipulations are usually fortuitous (Menn 1971) or applicable to only a few types of phenomena (e.g., Berko 1958).

Observation of the model will employ the same methods, indeed must employ these methods to establish whether its behavior is comparable with human behavior. Our basic methodology will be longitudinal observation and phonetic transcription of speech samples with appropriate statistical and structural analyses. We shall directly inspect internal representations and use experimental manipulations in ways that are not possible with real children.

Experimental manipulations are particularly important. In observing real children, fortuitous natural experiments occur at all times (e.g., introduction of new toys, people, and words in the child's environment), but the investigator rarely has control of such circumstances. We not only have control of these variables, but the entire social and physical environment in our model is synthetic. Thus, we must evaluate different synthetic environments if for no other reason than to introduce internal controls into the simulation design.

Some phenomena will naturally emerge without intervention. Other phenomena are not guaranteed to emerge as the child or synthetic system unfolds naturally. They may be accidents of the child's or model's history. We wish to test whether such behavior may be predicted from a particular history. For example, consonant harmony is not guaranteed; it seems to occur at statistically chance levels, and a systematic pattern of harmony is even less common (Vihman 1978). To investigate the origins of such behavior, we must hypothesize its cause and manipulate the model's history to test the hypothesis.

3.1.1. Automatic phonetic transcription

One advantage of our modeling environment is our ability to automatically transcribe and compile the model's babble and speech. In order to automatically transcribe the model's speech, we directly measure vowel quality from static spectra or formant frequencies. Alternatively, we train a neural network classifier to recognize parental demi-syllables and vowels or to replicate experimenter classification of model demi-syllables and vowels. We also use the model's own segment detection algorithm with an appropriate set of rules to analyze syllable and prosodic structure of utterances. We compile these and other analyses to generate a phonetic transcription and various structural analyses.

The validity of this automated transcription methodology must be tested against human observers trained in phonetic transcription. We must adjust or adapt the automated methodology to achieve cross-transcription agreement rates comparable to typical studies of child phonology (e.g., Vihman et al. 1986, Kent & Bauer 1985, Oller et al. 1975). Whether we can achieve this is an open question. The hand-crafted adult speech produced by ASY is rather unnatural because of limitations in ASY's voice and frication source model and other factors. The model's immature speech patterns may be even more difficult to transcribe.

3.2. Parental corpus

Two sources of data guide the model's development of perceptual categories and shape its future — parental speech and the model's own utterances. The model's own utterances are a function of constraints built into the vocal tract and gesture controller models and basic costs built into the reinforcement schedule. Parental speech can be subjected to considerable experimental manipulation, which we detail in sections below. Design of the basic corpus of parental speech will use one of the following two methods:

- **Vocabulary selection from a synthetically generated corpus.** We generate a database of one-syllable and simple two-syllable sounds which conforms to the phonotactic and syllable structure rules of English phonology. We select words from this database which (1) are statistically likely to occur in parent-child speech interactions, (2) represent stressed syllables of common polysyllabic utterances, (3) can be accurately pronounced by the articulatory synthesizer, and/or (4) represent a reasonable distribution of sound patterns from which the compositional structure of one-syllable utterances may be induced.
- **Vocabulary compilation from a vocabulary from actual samples of Motherese.** We compile this vocabulary from the CHILDES database (MacWhinney 1991), but filter out utterances which cannot be accurately pronounced by the articulatory synthesizer, and retain only the stressed syllable of polysyllabic words.

Once a basic vocabulary is transcribed, we convert each morpheme into a gesture script automatically or by hand. We check and adjust the pronunciation of each script. When realized as individual tokens of parental speech, gestural parameters will be perturbed randomly to introduce some statistical variation. The model will be exposed exclusively to parental speech for a period of time corresponding to the child's earliest experience.

3.3. Essential phenomena and properties

We hope to observe and wish to test for some subset of the following properties and types of phenomena: (1) specific statistical trends during babbling, (2) specific types of errors such as segment substitution and cluster reductions, (3) alternations among erroneous and accurate patterns (4) systematicity of errors, (5) overgeneralization of sound patterns, and (6) inertia of old sound patterns (Menn et al. 1993). We now consider these in greater detail.

3.3.1. Emergent behavior of model observed without experimental manipulations

The following phenomena and properties should emerge from the model's behavior without the need for experimental manipulations. However, some tests may require experimental controls or a population of model instances. We summarize or exemplify the data, explain how it is explained by the model, and propose an observational or experimental methodology.

- Children's babble increasingly approximates adult phonetics, statistically increasing in uniformity and correlation with the native language.

  Cross-linguistically, children's earliest babble shows a preference for labials, dentals, oral and nasal stops which does not necessarily match the statistics of adult speech (Oller et al. 1976, Kent & Bauer 1985, Vihman et al. 1986, de Boysson-Bardies & Vihman 1991). As time progresses, these trends give way to preferences more closely resembling parental models, showing less variance and higher correlations among subjects in the same linguistic community (Vihman et al. 1986, de Boysson-Bardies et al. 1992).

This behavior emerges from the model in the following way. The earliest vocalizations are mainly a function of random high-temperature flailing by articulators. Excess degrees of freedom are an advantage to the model at this time, enabling it to easily experiment with crude but adequate approximations of stop consonants in primitive syllables. The most probable vocalizations are precisely those observed in early babbling. However, reinforcement of phonological and articulatory controllers is based on the sound categories discovered in the linguistic environment by auditory perception. Rewards favor a distribution of produced sounds which resembles the distribution of perceived sounds.

The simplest test of this phenomena is a comparison of phonetic statistics for model and parental corpus, but it lacks any experimental control. We introduce a control by comparing at least two model instances, exposed to parental corpora with statistically different phonetic characteristics (e.g., distribution of vowels, stops, nasals, places of articulation). However, to reproduce variance and correlation results requires one or more populations of model instances, all individuals in each population sharing the same general linguistic environment but not sharing identical adult corpora. We measure vowel quality, vowel inventories, consonant inventories, syllable structure, and canonical syllable rates for both parental speech and model utterances.

- Children substitute one segment for another and delete segments from consonant clusters.

Children substitute stops for fricatives, glides for liquids, voiced consonants for word-initial voiceless consonants, and one place-of-articulation for another; they delete segments, especially from consonant clusters (e.g., Menn 1971, Smith 1973, Ingram 1974, reviews by Menn 1978, 1983, Ingram 1989).

The relative difficulty of pronouncing various sounds is not necessarily a function of effort or other costs. It is more closely related to relative combinatorial complexity and proprioceptive cues available to guide articulatory control. Stop consonants require the least complex motions and are aided by discrete cues; they also benefit from the model's excess degrees of freedom. Fricatives require more precise control but proprioceptive cues are less informative. In the early stages of phonological learning, there will be a higher probability of generating a phonetically similar if somewhat inaccurate syllable using a stop consonant rather than the target fricative. For example, "thank" [T&Nk] and "tank" [t&Nk] will have target phonetic state representations which agree in most respects; the correct pronunciation will show a static fricative segment where the latter shows an aspirated segment; onset spectral transitions will be very similar. Even if both pronunciations are represented among articulatory controllers, the phonological controller will choose the more likely, "tank". Similar arguments may be extended to other substitutions.

The simplest test of this phenomena requires a statistical compilation of the model's segmental substitutions or deletions at various stages of development.

- Errors are systematic, appearing regularly in the same phonological context.

Substitutions and deletions occur with rule-like regularity, and they usually occur in some limited context. As new words are introduced, they show the same error patterns. Examples include deletion of word-initial fricatives by some children (e.g., "fish" becomes [IS], "shoes" [uz], "soup" [up]), substitution of fricatives by stops ("soon" [dun], "fly" [baj]), voicing of all initial consonants, insertion of sounds in related contexts (e.g., insertion of palatal before /s/, "bus" becomes [bajs]), and many more.

The systematicity of the model's behavior derives from a phonetic representation which captures the similarity of neighboring sounds and the generalization properties of the phonological controller. The context sensitivity of its behavior emerges from the whole-syllable representation of target sounds, the correspondence between demi-syllables, articulatory gestures, and the representation of spectral transitions.

To test the systematicity and context sensitivity of segment substitutions and deletions requires conditional statistics, a set of rewrite rules which describe the errors and their contexts, or a set of canonical forms or prosodic structures which describe the regularities of utterances. Rewrite rules and canonical forms also reveal the underlying compositionality of utterances.

- Regular patterns of alternation occur between erroneous forms or between accurate and erroneous forms.

Regular error patterns change. Sometimes they give way to accurate, adult-like forms. But they often do so only gradually, following a period of competition between the old and new patterns. An interesting example is the alternation shown as Daniel's regressive velar harmony gave way to consonant contrast (Menn unpublished data, see Table 3). Note the alternation between the inaccurate /g/ onset and the correct forms with [b], [p], [d], and [t] in words with a final [g], [k], or [N].

Table 3: Daniel's Regressive Harmony As It Crumbles (Menn, unpublished data)

32;28	ai mEik A h&ws	"I make a house"
	gAg	"bug" (early in the day)
	bAg	"bug!" (alarmed by it; many repeats)
32;29	bIg	"big"
	baks	"box"
	b&g	"bag"
	tig&g	"teabag"
33;0	gaks / baks	"box"
	gEigw	"bagel"
	bEigw	"bagel" in response to L.'s model.
	uw gaks / uw baks	"little box"
33;1	dAn dAks	"done with the ducks" (elicited with effort)
	dAn gAks	"done with the ducks" (relapses)
	gAk / dAk	"duck" (spontaneous, correctable with relapses)
	ai pINkIN INz	"I pinking things" (looking through pink plastic)
	bivi tu gIg	"Stevie too big"
33;2	gUk / bUk	"book" (self-corrected)
33;3	dOg / gOg	"dog" (response to picture)
	pIg	"pig"
	wat u du^wIN	"what you doing?"
33;4	bIwd A gIg t&wr	"build a big tower"
	Is nat tu bIg	"this not too big?" (later, about spoonful of something)

For the simplest alternations, we hypothesize that articulatory controllers exist for each competing form and that the phonological controller is learning to invoke the right one. When component sounds compete, we argue that the Q-values associated with two alternative articulatory controllers are nearly equal and their activation is equiprobable at the point in the utterance when the phonological controller must choose between them.

Test methodologies are the same as those immediately above.

3.3.2. Model behavior which requires experimental manipulations or direct inspection to observe.

Testing the following properties and hypotheses requires experimental manipulation or direct inspection of internal model representations.

◆ Newly learned words show evidence of context sensitive generalization, but the oldest forms resist generalization to new sound patterns.

A new articulatory pattern is usually revealed in the introduction of a new word to the child's vocabulary. The new word might even be pronounced accurately. Whether accurate or not, it often spreads inappropriately to phonetically similar words. Daniel's nasal harmony rule started innocently enough in his pronunciation of "moon" [mun ~ mum], but then inappropriately spread to "broom" [mum], "mug" [NAN], "going" [NowIN], "spoon", and "prune" (various forms) (Menn 1971, unpublished data). Eventually, the lexical domain of this new error became more and more inclusive until even established forms for "down" [d&wn] and "stone" [don] became infected (e.g., [n&wn], [non]). This is a particularly virulent case of overgeneralization. It also reveals the relative resistance or "inertia" of older forms. Another example demonstrates a beneficial case of generalization, as well as inertia. Jacob substituted /d/ for /b/ in /b/-initial words until a relatively late age. His first accurate /b/-initial words were new additions to his lexicon. The new skill spread to other new /b/-initial words, but he took several weeks to correct older words, continuing to substitute /d/ for /b/ for a while.

We claim that overgeneralization is a consequence of the basic generalization properties of the phonological controller, the built-in generalization behavior of phonetic categorization, and the coarsely-coded phonetic state representation. We attribute inertia to memorization and overtraining of old forms. Recall that the phonological controller's phonetic goal is represented as a whole-morpheme pattern. Internal network representations are developed and reinforced which favor the old pronunciation even when competing, more accurate behavior first emerges. The model predicts that they may alternate until the old pattern is extinguished.

As a test methodology, we introduce new (natural or artificial) words which are similar to existing words in the model's productive vocabulary or which differ from the model's active vocabulary in certain dimensions.

◆ We hypothesize that compositional structure evolves from whole syllables to demi-syllables to segments.

Phonological structure shows a transition from whole-word to segmental organization as development proceeds (Macken 1979, Vihman & Velleman 1989). We believe that this behavior emerges as the phonological controller discovers that more target sounds may be produced more efficiently by recomposing parts of sounds previously mastered by articulatory controllers. The temporal granularity of the phonetic state and the recombinatorial constraints of articulatory controllers could allow a segmental phonological organization. (Here we use "segmental" in its traditional sense of phoneme-sized units of speech.)

Directly inspecting the behavior of the phonological controller should immediately reveal how it composes complex utterances, and how its behavior evolves.

◆ We hypothesize that consonant harmony depends on model history which biases its behavior.

The model contributes necessary but not sufficient conditions for consonant harmony (see section 3.3.2 and Table 3 for harmony examples). One component of the target's phonetic representation is word shape; its representation of spectral transition category does not explicitly state where in the word a demisyllable occurs. A "guck" matches many of the phonetic properties of a "duck", and likewise other harmonic forms satisfy most criteria for accuracy — word shape, vowel, and one demi-syllable. The ingredients of the basic model may be sufficient for infrequent occurrences of consonant harmony, but they are not sufficient for the systematic cases which have been documented (Menn 1971, Smith 1973, cf. Vihman 1976).

Several plausible phonological control tactics can construct a CVC syllable out of two CV or two VC syllables. If such a tactic is used, a previous preference for reduplicated babble may result in consonant harmony. Whether it lasts depends on the model's (or child's) tolerance for inaccurate pronunciation (Menn 1983).

To test this hypothesis, we introduce a preference for reduplicated consonants during babble. This will hopefully generate a bias for repeating consonants in CVC and multi-syllabic utterances in early speech. We control for tolerance by adjusting parameters of phonetic distance metrics.

3.3.3. Comparing exploration and bootstrapping strategies

In section 2.8.2. we propose two alternative bootstrapping strategies. In one, we artificially stage the complexity of target utterances. In the other, we allow emergent properties of the model and individual differences and preferences to pace the complexity of phonetic targets. The latter approach welcomes experimentation to determine the relative effectiveness of bootstrapping strategies, e.g., avoidance and selection based on predicted success, selection based on semantic and pragmatic criteria, or intrinsic sound preferences. Children employ all these techniques in very idiosyncratic ways (Menn 1983). Our experimentation may suggest questions for a more systematic investigation of exploration strategies and effectiveness.

3.3.4. Claims of model

The model claims that undisturbed auditory feedback is necessary for executing speech. In fact, slightly delayed auditory feedback severely disrupts normal speech (Lee 1950, Black 1951, Huggins 1964). The worst disruption occurs for delays equal to about the average length of a syllable. They cause stuttering, slurring, repeated syllables, and intense frustration. It also claims disruption of speech if proprioceptive feedback is disturbed. Evidence supporting this claim includes lip and jaw perturbation experiments (Kelso et al. 1984, Shaiman 1989) which affect the duration of voicing and studies of proprioceptive deficits (Abbs & Connor 1991).

We may further test these claims by artificially delaying or otherwise disrupting auditory feedback and by introducing noise, lesions to the proprioceptive state or simulating physical perturbations.

4. Model shortcomings and opportunities for future research.

Though it builds on a long tradition of research, this model is only a start, a crude approximation of the computational principles which may ground phonological development. There are several obvious shortcomings which to address is beyond the scope of this research. They do provide opportunities for future research.

We do not implement a phonological loop (Baddeley 1986, 1992). This precludes studying any phenomena which may require a phonological working memory, including observations that children with working memory deficiencies show deficits in the later stages of phonological development (Gathercole & Baddeley 1990). A phonological loop may be introduced by inserting an associative memory between the output of gesture controllers and early auditory analysis stages.

Proprioceptive reaction times in adults suggest a variety of mechanisms which short-circuit the explicit proprioceptive feedback we have built into the model (e.g., Evarts 1971, Taylor & Birmingham 1948, Higgens & Angel 1970). This does not invalidate the basic principles of the model with regard to the role of proprioception. But a more complete model would probably introduce various predictive components which anticipate proprioceptive (or phonetic) feedback and allow planning behavior beyond the scope of the present model.

Adult behavior implies that the organization of phonological representation is based not on phonemic segments but rather on a system of distinctive features (Chomsky & Halle 1968, Durand 1990). Such features are grounded in phonetic differences but signify semantic distinctions. Our model does not address this. It only addresses the phonetic and articulatory foundations of a more mature phonology.

The articulatory synthesizer has a number of weaknesses which limit the utterances which the model may attempt (section 2.10.). Its greatest weakness is its inability to reconfigure the dimensions of the vocal tract to conform to a child's anatomy (e.g., Goldstein 1980). Thus, our model cannot address the difficult problem of perceptual vocal tract normalization, which may be an interesting factor in early phonological development.

We do not grant any role to attentional processes. This may be crucial to bootstrapping. It may be crucial to many phenomena including consonant harmony. Nor do we acknowledge any communication between phonological and articulatory controllers other than activation and distribution of reinforcement. This may not be the most efficient solution, vastly increasing the number of articulatory controllers which may be needed for some target language. A more efficient solution may pass certain limited signals between the two levels without violating Markov decision task constraints.

5. Research Plan

Our work to date has implemented and demonstrated the parallel Q architecture (Markey 1994) and the successful operation of articulatory controllers operating without the phonological controller to generate some crude CV syllables (see figure 5 and accompanying text, section 2.7.3.). We have not yet integrated auditory perception with articulation, so the utterances are not quite natural. But they are sufficient to demonstrate the concepts. We have also implemented most of the components of the articulatory control loop, including the gesture controllers, proprioceptive analysis, and articulatory synthesizer.

We have also implemented and tested much of the auditory perception machinery, including spectral analysis and segment detection (Markey & Bell 1993), but most of prosodic analysis and all categorization modules remain to be built.

Also remaining are short-term, long-term, and lexical memories, the reinforcement function, a slightly more realistic voice and frication source model for the articulatory synthesizer, and input representations for the phonological controller. Some additional sensory features, event detectors, and an improved representation of of proprioceptive state may improve performance of the articulatory controllers. Most critically, we must integrate auditory perception with phonological control. We must generate or automate production of the parental corpus. Finally, we must implement automatic phonetic transcription and a record-keeping procedure. I estimate that we have completed about 60% to 70% of the code necessary for the project.

Initial simulations will be devoted to some small parental corpus of demi-syllables and then adds target words composed of the mastered sounds. They will also be devoted to model calibration and parameter tuning. Once the model's basic behavior is confirmed, we may start more elaborate simulations and experimental manipulations.

I intend to have sufficient early results in time for Cognitive Science Society and Machine Learning submissions, with more complete results and analysis in time for a NIPS submission. I intend to complete my dissertation by the end of Summer 1994.

Acknowledgements

This research is supported by NSF Presidential Young Investigator award IRI-9058450 and grant 90-12 from the James S. McDonnell Foundation to Michael C. Mozer. We thank Haskins Laboratories for making available its ASY software and source code, and Troy Sandblom for his rewriting it for C and Unix. I acknowledge the tireless guidance and coaching by Mike Mozer and Lise Menn — Mike for his focus, insight, and willingness to endure this, Lise for her vision and for piloting me through the phenomena. Thanks are due other committee members: Clayton Lewis for asking the question in the first place, Paul Smolensky for inspiration and high standards, Alan Bell for answering my many signal processing questions, Jim Martin for perspective, and all for discussions and feedback. I also wish to thank the following people (and the several whom I may have overlooked) for many discussions of their work or the present research: Chuck Anderson, Peter Dayan, Patricia Kuhl, Rafael Laboissiere, George Lakoff, Brian MacWhinney, Steve Nowlan, Satinder Singh, Gerry Stahl, Rich Sutton, Sebastian Thrun, Richard Yee, Janet Werker. To members of the Boulder Connectionist Research Group also go thanks, especially Brian Bonnlander and Robert Dodier for discussions of the parallel Q architecture. Finally I express my deepest gratitude for Candice Miller's love, vision, and support, without which this work would not have been possible.

References

Abbs, J.H. and Connor, N.P. (1991). Motorsensory mechanisms of speech motor timing and coordination. *Journal of Phonetics 19:* 333-342.

Baddeley, A. (1992). Working memory. Science 255: 556-559.

Baddeley, A. (1986). *Working Memory.* New York: Oxford University Press.

Benedict, H. (1979). Early lexical development: comprehension and production. *Journal of Child Language 6:*183-200.

Berko, J. (1958). The child's learning of English morphology. *Word 14:* 150-177.

Black, J.W. (1951). The effect of delayed side-tone upon vocal rate and intensity. *Journal of Speech and Hearing Disorders 16:* 56-60.

Bradshaw, G. and Bell, A. (1991). Robust feature detectors for speech. Cognitive Science Technical Report UIUC-BI-CS-91-17. Urbana, IL: Unversity of Illinois, The Beckman Institute.

Browman, C. P. and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology 6:* 102-151.

Carpenter, G.A. and Grossberg, S. (1987). ART2: self-organization of stable category recognition codes for analog input patterns. *Applied Optics 26:* 4919-4930.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English.* New York: Harper and Row.

Daugherty, K. and Seidenberg, M.S. (1992). Rules or connections? The past tense revisited. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society,* pp. 259-264. Hillsdale, NJ: Erlbaum.