THE SENSORIMOTOR FOUNDATIONS OF PHONOLOGY: A COMPUTATIONAL MODEL OF EARLY CHILDHOOD ARTICULATORY AND PHONETIC DEVELOPMENT

Kevin L. Markey

CU-CS-752-94 1994

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

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the agencies named in the acknowledgments section.

The Sensorimotor Foundations of Phonology:

A Computational Model of Early Childhood Articulatory and Phonetic Development

by

Kevin Lee Markey

B.A., University of Chicago, 1988
M.S., University of Colorado, 1991

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Computer Science

December 1994
Revised February 1995

Copyright © 1994, 1995 by Kevin L. Markey

Language, whether spoken or signed, partly defines us as human beings. This thesis is for the children who by accident, force, or neglect lose or never find their voice — with a prayer that they reclaim it.

And
for Candice.

Abstract

This thesis describes HABLAR, a computational model of the sensorimotor foundations of early childhood phonological development. HABLAR (an acronym for "Hierarchical Articulatory Based Language Acquisition by Reinforcement learning" and Spanish for "to speak") is intended to replicate the major milestones of emerging speech and demonstrate key characteristics of normal development, including the phonetic characteristics of babble, systematic and context-sensitive patterns of sound substitutions and deletions, overgeneralization errors, and the emergence of adult phonemic organization. It should also mimic abnormal phonological development under certain conditions of damage or degradation.

HABLAR simulates a complete sensorimotor system consisting of an auditory system that detects and categorizes speech sounds using only acoustic cues drawn from its linguistic environment, an articulatory system that generates synthetic speech based on a realistic computer model of the vocal tract, and a hierarchical cognitive architecture that bridges the two. The environment in which the model resides is also simulated. The model is an autonomous agent which actively experiments within this environment.

The principal hypothesis guiding the model is that phonological development emerges from the interaction of auditory perception and hierarchical motor control. The model's auditory perception is specialized to segment and categorize acoustic signals into discrete phonetic events which closely correspond to discrete sets of functionally coordinated gestures learned by the model's articulatory control apparatus. HABLAR learns the correspondence between phonetic and articulatory events — between one discrete sequence of events and another. Thus, to imitate sounds, it need not solve the hard problem of relating continuous speech and continuous vocal tract motion.

HABLAR's perceptual and motor organization is initially syllabic. Phonemes are not built into the model but emerge (along with an adult-like phonological organization) due to the differentiation of early syllable-sized motor patterns into phoneme-sized patterns while the model learns a large lexicon.

Learning occurs in two phases. In the first phase, HABLAR's auditory perception employs soft competitive learning to acquire phonetic features which categorize the spectral properties of utterances in the linguistic environment. In the second phase, reinforcement based on the phonetic proximity of target and actual utterances guides learning by the model's two levels of motor control. The phonological control level uses Q-learning to learn an optimal policy linking phonetic and articulatory events. The articulatory control level employs a parallel Q-learning architecture to learn a policy which controls the lung's and vocal tract's twelve degrees-of-freedom.

HABLAR has been fully implemented as a computational model. Simulations of the model's auditory perception demonstrate that it faithfully preserves and makes explicit phonetic properties of the acoustic signal. Auditory simulations also mimic categorical vowel and consonant perception which develops in human infancy. Other results demonstrate the feasibility of learning multi-dimensional articulatory control with a parallel reinforcement learning architecture, and the effectiveness of shaping motor control with reinforcement based on the phonetic proximity of target and actual utterances.

The model also provides plausible qualitative accounts of developmental data. It is predicted to make pronunciation errors similar to those observed among children because of the relative articulatory difficulty of its producing different speech sounds, its tendency to eliminate the biggest phonetic errors first, its generalization of already mastered sounds across phonetic similarities, and contextual effects of phonetic representations and internal distributed representations which underlie speech production.

Contents

List of Figures ix

List of Tables x

Acknowledgments xi

Chapter 1 Introduction 1

1.1 Milestones of normal phonological development 1

1.2 Approaches to investigating phonological development 2

1.3 A sensorimotor model of phonological development 2

1.4 The model's principal hypothesis 3

1.5 Related models and the unique contributions of HABLAR 3

1.6 Plan of the thesis 4

Chapter 2 Empirical Constraints on Models of Phonological Development 5

2.1 Pronunciation errors and mastery 5

2.2 Generativity and context-sensitivity 5

2.3 Relationships between adult target words and children's speech 6

2.3.1 Implications of realization rules 7

2.4 Syllabic and whole-word constraints 8

2.5 Overgeneralization and inertia 9

2.6 Irregularities 10

2.7 Variability 10

2.8 Perception of speech and underlying representations 11

2.8.1 Representation of target utterances 12

2.8.2 Phonemic vs. phonetic perception 12

2.8.3 Syllabic and dynamic phonetic perception 13

2.8.4 Model constraints based on speech perception data 14

Chapter 3 Computational Models of Phonological and Articulatory Development 15

3.1 Speech perception models 15

3.2 Two classes of speech production models 15

3.3 Abstract phonological competence models 16

3.4 Articulatory skill models 17

3.4.1 Kinematics of articulatory and acoustic trajectories 17

3.4.2 Forward model error inversion and articulatory skill 18

3.4.3 Dynamics of articulatory skill acquisition 18

3.4.4 Inverse dynamics with temporal credit assignment problems. 19
3.4.5 Dynamic components of articulatory skill models 20
3.5 Reinforcement learning as an alternative control strategy 21
3.5.1 Advantages and disadvantages of reinforcement learning 22
3.5.2 Models of hierarchical reinforcement learning and control 22

Chapter 4 HABLAR's Architecture and Behavior 24
4.1 Empirical constraints and problems of speech production. 24
4.2 Resolving issues of phonological development 24
4.3 Model components 25
4.4 Imitating ***duck***: A sample scenario. 27
4.5 The model's operation 28
4.6 Components that learn 29
4.7 Phased learning. 29
4.8 Built-in knowledge and components 30

Chapter 5 Articulatory Anchors and Speech Generation 31
5.1 Speech generation by the articulatory synthesizer and source model 31
5.2 Gestural organization of articulatory motions 32
5.3 Proprioceptive perception. 33
5.4 Source model implementation details 34

Chapter 6 Auditory Perception 37
6.1 Motivation and Background. 37
6.2 An acoustic-based phonetic representation 38
6.2.1 Form of the phonetic representation 39
6.2.2 Spectral features 40
6.2.3 Segmental features 40
6.2.4 Superimposed activations and phonetic distance 40
6.2.5 Phonetic categorization and spectral category learning 41
6.2.6 Redundancy between noncontextual and contextual features. 41
6.3 Auditory perception implementation details 42
6.3.1 Computational resources 42
6.3.2 Spectral analysis 43
6.3.3 Learning spectral features 45
6.4 Auditory Perception Simulations 46
6.4.1 Stimuli 46
6.4.2 Relative phonetic distance as a measure of distinctiveness 47
6.4.3 Supervised classification as a test of faithfulness 48

6.4.4 Locality and faithfulness of categorization 48
6.4.5 Emulating the acquisition of categorical perception 49

Chapter 7 Hierarchical Motor Control 51
7.1 Empirical constraints governing motor control 51
7.1.1 Phonetic vs. articulatory and linguistic events 51
7.1.2 Coordinative structures and the partition of motor control 51
7.2 Form and function of hierarchical motor control 53
7.2.1 Formalizing control problems as Markov decision tasks 53
7.2.2 Articulatory controllers 54
7.2.3 Phonological controller 56
7.3 Relationships between control levels 56
7.3.1 Role of phonological controller in linguistic composition 56
7.3.2 Role of the articulatory controller in linguistic composition 57
7.3.3 Ideal operation illustrated 58
7.3.4 How many articulatory controllers? 60
7.4 Computational issues of control 60
7.4.1 Why reinforcement learning? 60
7.4.2 Continuous-valued state variables 61
7.4.3 Function approximation and aliasing 61
7.4.4 Ensuring Markovian rewards 61
7.4.5 State isolation and reward hiding 62
7.4.6 Nonstationary search and hierarchical search complexity 63
7.4.7 Bootstrapping the hierarchy 63
7.5 Controlling HABLAR's multi-dimensional action space 64
7.6 Controller implementation 65
7.6.1 Neural network implementation of controllers 66
7.6.2 Reinforcement functions 66
7.6.3 Phonological controller implementation 67
7.6.4 Articulatory controller implementation 67
7.7 Motor control simulations 68
7.8 Why does parallel Q-learning work? 68
7.8.1 Cooperation emerging from self-interested behavior 69
7.8.2 Limitations of parallel-Q's sparse representation of action 71
7.8.3 Nonstationary environments and parallel subagents 72
7.8.4 Empirical success 72

Chapter 8 Results of Motor Control and Integrated Model Simulations 73
8.1 Proximal task simulations. 73
8.1.1 Stop-consonant motion 73
8.1.2 Primitive consonant-vowel syllables. “Deaf” babble. 75
8.1.3 Reduplicated babble 75
8.1.4 Training failures 76
8.2 Integration of audition and motor control 76
8.2.1 Procedure. 77
8.2.2 Results 77
8.3 An engineered hierarchy. 79
8.4 Phased complexity simulations 79

Chapter 9 Explaining Developmental Psycholinguistic Data 80
9.1 HABLAR’s properties which contribute to observed phenomena. 80
9.1.1 Effects of articulatory mechanisms 80
9.1.2 Effects of stochastic gradient descent learning algorithms 81
9.1.3 Effects of the phonological controller’s associative architecture 81
9.1.4 Effects of external and internal phonetic features 81
9.1.5 Effects of the phonological controller’s model of speech sounds. 82
9.1.6 Other effects of computational mechanisms 83
9.2 Qualitative Explanations and Predictions. 83
9.2.1 The phonetic characteristics of babble 83
9.2.2 Sound substitution 84
9.2.3 Sound competition. 85
9.2.4 Voicing contrast 85
9.2.5 Context sensitive generalization 86
9.2.6 Consonant harmony. 86
9.2.7 Pronunciation difficulties and learning strategies 90
9.2.8 Emergence of phonological awareness 91
9.3 Explanatory shortcomings due to HABLAR’s scope or design 93
9.3.1 Coarticulation and assimilation. 93
9.3.2 Multisyllabic targets and utterances 94
9.3.3 Rapid adult speech. 94
9.3.4 Articulatory limitations 95
9.4 HABLAR and phonetic speech perception 95
9.4.1 Motor and direct perception theories of speech perception 95
9.4.2 Grounding the notion of direct perception 96

Chapter 10 Conclusions and Future Directions 98
10.1 Design for an empirical test of HABLAR's fidelity 98
10.1.1 The phonetic characteristics of babble 99
10.1.2 Sound substitution and deletion 99
10.1.3 Context sensitive systematicity of pronunciation errors 99
10.1.4 Overgeneralization and inertia 99
10.1.5 Articulatory precision and bootstrapping strategies 100
10.1.6 Emergence of phonological segmentation and awareness 100
10.2 Candidate model extensions and applications 100
10.3 Contributions 101
10.3.1 Technical features 101
10.3.2 Systemic constraints on neural network hypotheses 101
10.3.3 A new testbed and analytic framework 102

Bibliography 103

List of Figures

Figure 3.1 Components that define a Markov decision task 21
Figure 4.1 HABLAR's cognitive architecture....26
Figure 4.2 Acoustic signal and phonetic events....27
Figure 5.1 Vocal tract articulators 31
Figure 5.2 Articulatory gestures....33
Figure 6.1 Idealized spectrogram of ***duck*** before and after segmentation 39
Figure 6.2 Modeling clusters of observations as 2-dimensional Gaussians....41
Figure 6.3 Information flow in auditory perception....43
Figure 6.4 Static and transition spectra for onset demisyllable of ***duck*** 44
Figure 6.5 Transition magnitude for duration of ***duck*** 45
Figure 7.1 Articulatory gestures vs. spectral transitions 52
Figure 7.2 Nested phonological and articulatory control levels....57
Figure 7.3 Intended HABLAR behavior....59
Figure 7.4 Standard Q-agent architecture 64
Figure 7.5 Quasi-independent subgent architecture....65
Figure 7.6 Value functions for standard and parallel Q-learning 70
Figure 8.1 Articulatory controller learning consonant-like motion 74
Figure 9.1 A statistical mixture model of the linguistic environment....83
Figure 9.2 Sound substitution....85
Figure 9.3 Consonant harmony — how ***duck*** becomes ***guck*** 87
Figure 9.4 Place-of-articulation variation in a complex morpheme....91

List of Tables

Table 2.2 Labial prosodic structure 9
Table 2.3 Crosstalk among several of Jacob's related words 10
Table 2.4 Variation in Daniel's regressive harmony as it crumbles....11
Table 6.1 Segment types and defining acoustic cues....40
Table 6.2 Mean phonetic distance measured between bV syllable tokens....47
Table 6.3 Additional phonetic distance comparisons 47
Table 6.4 Supervised training error rates....48
Table 6.5 Unsupervised categorization of static spectra....49
Table 7.1 Acoustic segments vs. articulatory and linguistic events 52
Table 7.2 Proprioceptive state and articulatory controller inputs....55
Table 7.3 Summary of gestural targets 55
Table 7.4 Idealized articulatory sequence which generates /ba/....62
Table 7.5 Articulatory sequence which does not generate /ba/ 62
Table 8.1 Stop-consonant contact and release 75
Table 8.2 Articulatory controller learning of primitive syllables 75
Table 8.3 Select integrated sensorimotor simulations 78
Table 9.1 Infant stop consonants vs. HABLAR vocal tract constrictions 84
Table 9.2 Phonetic distance between target and various pronunciations....88

Acknowledgments

If there is one thing I will prize about the four years spent on this project, it will be Mike Mozer's generosity — with his time, coaching, insights, feedback, encouragement, confidence, credit, and humor. Whether suggesting a new viewpoint when I was unable to solve a problem, focusing me with a key question when I was lost in a sea of details, steering me away from an unproductive tangent, or editing yet one more revision, Mike was always available. Moreover, he allowed me to cause trouble when I really needed to. Mike lost patience only when I held him hostage as I fished for the right word or idea.

Most importantly, I appreciate Mike's willingness to endure this. There have been times when this project looked hopeless, but Mike kept faith and good humor, always forthcoming in intellectual and material support. (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).

I also acknowledge Lise Menn's guidance and support. I am thankful for her vision, her enthusiasm, and her help in navigating the vast sea of language acquisition literature in search of phonological phenomena. This project would not have been half as rich without her broad vision of child phonology, nor half as fun.

Thanks are due other committee members. Clayton Lewis asked the question of Lise which prompted the project. I appreciate their willingness to have me hazard a guess on its answer. Paul Smolensky was key to my returning to graduate school. I thank him for the inspiration of his own work and the high standards he set for mine. Thanks go to Alan Bell for answering my questions about acoustic phonetics and to Jim Martin for perspective and a sympathetic ear, even at short notice when I appeared at his always-open door. To all, I appreciate the moral support, discussions, and feedback.

Haskins Laboratories made available its ASY software and source code, and Haskins' Philip Rubin patiently answered questions about the code. I also owe researchers at Haskins a huge intellectual debt (apparent in citations) for their pioneering work in understanding speech perception and production.

Troy Sandblom rewrote ASY for C and Unix. Without this resource, the project would have been impossible. His contribution has paid many dividends. Dave Wood and Rich Harrison contributed software. CSOps kept the lab's computers in good health. Thanks to them, Karl Winklmann, and Dotty Foerst.

I thank the following people (and the several whom I may have overlooked) for the insights they contributed in discussions of their work or the present research: Chuck Anderson, Andy Barto, Peter Dayan, Prahlad Gupta, Michael Jordan, Peter Jusczyk, Rafael Laboissiere, Long Ji-Lin, Brian MacWhinney, Morgan Nelson, Steve Nowlan, Bruce Pennington, Jim Sawusch, Satinder Singh, Gerry Stahl, Rich Sutton, David Touretzky, Sebastian Thrun, Guy Van Orden, Richard Yee, Janet Werker, plus many visitors to CU or the lab who occasionally got an earful and lent moral support, pointers, or ideas to the project. To members of the Boulder Connectionist Research Group also go thanks, especially Brian Bonnlander and Robert Dodier for discussions of the parallel Q architecture, Bruce Tesar for help in understanding some concepts of probability theory, Sreerupa Das for debugging code and concepts, Franco Callari for software engineering assistance, and Clark Fagot for his comments on a draft of the thesis. Michael Carry, Course Director for Human Anatomy, CU Medical School, graciously allowed me to participate in the anatomy course for a few days.

Two people who may seem somewhat remote from the immediate project contributed more than they might imagine. Twenty-four years ago at the University of Chicago, Professor David McNeill and I argued whether it might be possible to build a rigorous sensorimotor theory of language acquisition. That dream was revived by George Lakoff, who in a meeting seven years ago introduced me to connectionism and indirectly to Paul Smolensky.

A big hug goes to my Mom for her support and the example of her life-long love of learning, a prayer of thanksgiving to my Dad for his exacting standards and integrity, tribute to my kid sister, Associate Professor Karen Drabenstott, for the friendly competition, and thanks to Aunt Ceil, in-laws Alan and Darlene and nieces Petra and Jana for their love and encouragement. Appreciation is due friends and clients for tolerating me, encouraging me, and keeping me and Candice humored through this process: Bill, Carolyn, the Six-Day Crew, Masselli, Roach, Geoff, Brant and Susan, Eileen, Bettina, Vim and Estella, Holly and Jock, Ted and Cindy, Gerry and Carol, Dan and Nancy, Ken and Kathy, Janet, Doug, and Sigrid.

Finally I express my deepest gratitude for Candice Miller's love, vision, humor, and patience, without which this work would not have been possible.

Chapter 1 Introduction

Auntie! Squirrels do not hop-hop!
'&nti: 'skwAwoz du 'na:t hap hap[1]

The words of children fill their parents, aunties, and uncles with joy, wonder, pride, and often laughter. Even as they grasp the fundamentals of syntax and master the intricacies of comparative animal behavior — after all, rabbits hop, not squirrels — they continue to struggle with how to pronounce the sounds of their language.

1.1 Milestones of normal phonological development

Children's linguistic journey starts early. Even before they can speak, they learn what speech sounds are linguistically important. At birth children distinguish acoustically different sounds whether they differ linguistically or not. But, by their sixth month, infants show a preference for prosodic patterns of pitch, stress, and duration used in their native language (e.g., Jusczyk 1992). They also tend to ignore acoustic differences among vowels of the same type in experiments where sound differences are used to elicit some response such as turning their head to view a toy. They do not group together foreign language vowel sounds (Kuhl et al. 1992, see also Grieser & Kuhl 1989). By eight months, the tendency to group together linguistically similar sounds extends to consonant-vowel syllables used in their native language (e.g., Werker et al. 1981, Werker & Pegg 1992). By the end of their first year, children can recognize about fifty words (Benedict 1979), and their receptive vocabulary quickly expands.

Milestones of speech production lag behind speech perception. Anatomical constraints prevent true speech sounds before three months (Kent & Murray 1982), but by the age of six months children's babbling starts to show characteristics of adult speech (Oller & Lynch 1992). By the end of the first year, their babbling approximates many of the sounds in their native tongue (de Boysson-Bardies et al. 1989, Vihman et al. 1986), and they speak their first real words (Benedict, 1979). In the second year, their vocabulary mushrooms to several hundred words.

Despite this impressive achievement, children do not immediately master accurate pronunciation. Their speech is filled with errors of commission and omission, only roughly approximating adult speech (e.g., Ingram 1974, Menn 1983). 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 (Menn 1971, Smith 1973). Instead, error patterns suggest an underlying competence, and changes in error patterns over time suggest that the underlying competence is modified by experience.

---

[1] Due to typographic limitations, UNIBET (MacWhinney, 1991) instead of IPA phonetic symbols are used throughout this thesis. UNIBET symbols which differ from IPA's include: N as in pi*ng*, T e*th*er, D ei*th*er, S *sh*oe, Z a*z*ure, I b*i*t, E b*e*t, & b*a*t, A b*u*t, 6 *a*bove, U f*oo*t, O l*aw*.

1.2 Approaches to investigating phonological development

Children's evolving mastery of their language's sounds and the changing patterns of their errors is traditionally called "phonological development" (Jakobson 1941/1968) despite its substantially phonetic and articulatory character. Indeed, this is not "phonology" in the usual sense. Adult phonology involves the study of distinctive classes of sounds (phonemes) and their possible combinations (phonotactics).

In investigating child phonology one may study the origin and evolution of pronunciation patterns and errors, the articulatory and phonetic foundations of phonological patterns, or how representations postulated to underlie adult speech patterns come to be. Most research has focussed on error patterns. Several theories and hypotheses have emerged to explain this phenomenon (e.g., Kiparsky & Menn 1977, Menn 1983, Stampe 1969), and several formalisms have been used to better characterize it (e.g., Smith 1973, Spencer 1986, Waterson 1971). The principal finding of such work is that there is a more-or-less regular relationship between children's and adult word forms, in which children simplify adult targets with substitutions, deletions, or other changes.

Yet, persistent irregularities (Menn & Matthei 1992, Menn et al. 1993) and idiosyncratic development strategies (Vihman 1993) frustrate a completely formal account. This is not surprising. Children's speech is not organized around phoneme-sized segments (Jusczyk 1993) which form the basis of linguistic formalisms. Rather, their speech production is organized around articulatory gestures (Goodell & Studdert-Kennedy 1993, Nittrouer et al. 1989, Piroli 1991), and their speech perception is organized around the syllable (Jusczyk et al. 1995, Segui et al. 1990). Only gradually, after having acquiring a large vocabulary, is children's speech reorganized into units the size of phonemes (Nittrouer et al. 1989). Given the degrees of freedom which govern phonology, diverse environments in which children learn to speak, and the accidents of each child's history, developmental paths are bound to be diverse.

A process model may be more successful than a formal symbolic model in explaining basic phenomena, irregularities, and individual differences. In traditional linguistic terms (Saussure 1916/1966), this would roughly correspond to a diachronic account, one which describes how linguistic structure changes historically or developmentally (McNeill 1987). In contrast, a synchronic account describes linguistic structure only at an instant in time. A longitudinal account of language development is usually viewed as diachronic. But at its heart, such a view of change as simply a series of instantaneous snapshots is synchronic. We seek an explanatory theory of phonological development: not merely the description of a sequence of changes in linguistic structure, but the structure and process of linguistic change. Central to such an account must be a learning theory which offers a principled explanation of how competence evolves. Another goal is to see how much we can explain without linguistic assumptions.

1.3 A sensorimotor model of phonological development

This work explores the articulatory and phonetic foundations of children's phonology and investigates the processes which underlie children's articulatory achievements and errors. It describes HABLAR, a computational model of the sensorimotor foundations of early childhood phonological development. HABLAR (an acronym for "Hierarchical Articulatory Based Language Acquisition by Reinforcement learning" and Spanish for "to speak") features (1) an auditory system that detects and categorizes speech sounds using only acoustic cues drawn from the model's own synthetic speech or from its linguistic environment, (2) an articulatory system that generates synthetic speech based on realistic computer models of respiratory mechanics, vocal tract anatomy, acoustics, and muscular-skeletal dynamics, and (3) a hierarchical cognitive and motor control architecture that bridges the two. The environment in which the model resides is also simulated.

HABLAR is intended to mimic how infants and young children babble and learn to speak their first words. We also hope to explain key characteristics of normal phonological development such as (1) the growing conformity of babbling sounds to the linguistic environment, (2) patterns of sound substitutions and deletions, (3) the systematicity of pronunciation errors, (4) the overgeneralization of pronunciation patterns, (5) the increasing ability to productively recombine smaller units of sounds, and (6) developmental irregularities.

1.4 The model's principal hypothesis

The principal hypothesis guiding the model is that phonological development emerges from the interaction of auditory perception and hierarchical motor control. The model's auditory perception is specialized to segment and categorize acoustic feedback into discrete phonetic events which closely correspond to discrete sets of functionally coordinated articulatory gestures learned by the vocal tract's motor control apparatus. The model learns the correspondence between discrete sequences of these phonetic and articulatory events. Thus, to imitate sounds, it need not solve the hard problem of relating continuous speech and continuous vocal tract motion. An adult-like phonological organization emerges as a consequence of this interaction while learning a large lexicon, due to the differentiation of early syllable-sized motor patterns into phoneme-sized motor patterns. The phenomena of phonological development are emergent properties of the model.

1.5 Related models and the unique contributions of HABLAR

Two related classes of computational models predate HABLAR: models of articulatory skill acquisition (Guenther 1994, Hirayama et al. 1993, Laboissiere et al. 1990, Laboissiere 1992) and models of abstract phonological competence (e.g., Daugherty & Seidenberg 1992, Gasser 1993, Ling & Marinov 1993, MacWhinney 1978, MacWhinney et al. 1989, MacWhinney & Leinbach 1991, Plunkett & Marchman 1991, Rumelhart & McClelland 1986). Models of articulatory skill acquisition address details of vocal tract control and the coarticulatory side effects of recombining sounds but not the origins or process of sound compositionality. Models of abstract phonological competence address compositional issues and the representational forms which underlie the recombination of sounds without acknowledging articulatory details or the origins of phonological or phonetic primitives used to represent sounds.

To address both the relative articulatory difficulty of different sounds and the recombination of already mastered sounds into unique utterances, HABLAR views speech as a hierarchical motor control problem. An abstract phonological level of motor control composes each utterance out of one or more elemental sounds, choosing which among lower level articulatory controllers is most likely to generate each sound. With continuous proprioceptive feedback as a guide, the articulatory control level choreographs the fine-grained timing of vocal tract and pulmonary motions necessary to produce each elemental sound. HABLAR also addresses computational problems of hierarchical control and how to efficiently control the vocal tract's many degrees-of-freedom.

The model's segmentation and categorization of acoustic feedback into discrete phonetic events is based on exclusively acoustic cues. Milestones of perceptual development always precede corresponding milestones of motor development. Thus, HABLAR's motor learning is preceded by a phase of perceptual learning in which phonetic primitives used to represent speech are learned from their statistical distribution in the model's linguistic environment (see also Doya & Sejnowski 1994).

The model avoids reference to any explicitly linguistic knowledge. If we give the model phonemes, its task may be too easy. Prior knowledge exists in the form of architectural, anatomical, and acous-

tic constraints, the form of prelinguistic phonetic representations, motor dynamics, learning algorithms, and reinforcement functions.

The model may not entirely satisfy parents' and relatives' questions about their children's intriguing chatter. However, the model and computer simulations of its behavior are potentially useful to the student of language acquisition in several ways. The precision required of computer simulation requires that theory and assumptions be explicit (Weizenbaum 1976). Its account of cognitive development is scientifically verifiable (MacWhinney 1978). It offers a vehicle with which to evaluate and test alternative assumptions. Also, simulations may reveal emergent properties which mere thought experiments or verbal argumentation might never uncover.

1.6 Plan of the thesis

The next chapter considers the psycholinguistic evidence which constrains the design of a model of phonological development. Chapter 3 considers the models which predate HABLAR in light of these issues. Chapter 4 introduces HABLAR, summarizing its architecture and basic behavior using a simple example, and discussing how its design addresses theoretical issues.

Anchoring and constraining the model at one end are its vocal tract, synthetic acoustic output generated by parent and child, proprioception, and gestural control of articulatory motions. These are described in Chapter 5. Anchoring the model at the other end is auditory perception. Chapter 6 describes how audition segments continuous speech using exclusively acoustic cues, how it learns to categorize these sounds, and how the resulting phonetic representation is organized. Simulations demonstrating the faithfulness of this representation are described. The methodology for generating synthetic adult consonant-vowel syllables is also described here. These syllables are used to test the faithfulness of auditory perception. They become part of the lexicon or speech tokens which the model learns to imitate.

Motor control is described in Chapter 7, including details of phonological and articulatory controllers, the reinforcement learning algorithms which they employ, and the role of the phonological controller in integrating the full sensorimotor model. HABLAR's control structure is formalized. Also, a parallel reinforcement learning architecture necessary for efficient control of the vocal tract is introduced and analyzed. Chapter 8 reports results of articulatory motor control and results of initial simulations of the integrated sensorimotor system.

At this time it is not possible to draw conclusions about how faithfully the system replicates developmental phenomena, but it is possible to evaluate how plausible are HABLAR's qualitative explanations of such phenomena. Chapter 9 evaluates the model's plausibility and limitations, considering a wide range of phenomena and properties of normal childhood phonological development. The final chapter explores the contributions and future directions of this research.

Chapter 2 Empirical Constraints on Models of Phonological Development

The literature on phonological development is vast. Menn (1983) and Ingram (1989: chapters 5, 6, and 8) review much of the early work in the field. Recent research is covered by Gerken (1994) or by Vihman's (1993) survey of individual differences in language-learning strategies.

It is not our goal to recapitulate these reviews. Rather, the purpose of this chapter is to abstract key properties of children's speech production and perception which are crucial to formulate and constrain a model of human phonological development. We ignore many of the details and several of the lesser issues of phonological development. Otherwise, the task of building a model will be unmanageable, and its implementation may be impractical or uninterpretable. Once the model is built and its performance has been evaluated with respect to the phonological data which motivated it and that which did not, then it may be refined (or replaced) to account for issues and details initially ignored in its construction. If we formulate the right abstractions and choose the right model, we may be fortunate enough to see phenomena explained or predicted independent of their role in model design.

Thus, in this chapter we review the relevant phenomena of children's speech and listening — pronunciation errors, generativity, context sensitivity of errors, the nearly systematic relationship between children's speech and adult targets, overgeneralization, irregularities, variability, and perceptual capabilities — and determine how they constrain a model of phonological development.

2.1 Pronunciation errors and mastery

Children's speech only roughly approximates adult speech. Perhaps the most characteristic pronunciation error is children's substitution of one sound for another: stops for fricatives, glides for liquids, voiced consonants for word-initial voiceless consonants, and one place-of-articulation for another. Children also delete sounds, especially from consonant clusters (Ingram 1974, Menn 1978, 1983, Smith 1973). Table 2.1 presents examples (Macken 1979, Menn 1971, 1976) of some common pronunciation errors.

Although specific error patterns are unique to each child, the universal occurrence of most types of errors and the approximate order with which children master various types of sounds suggest that some sounds are easier to pronounce than others. Most stop consonants and nasals are mastered first, followed by glides, then liquids, and finally fricatives (Kent 1992, Sander 1972; see also Ingram 1989). The articulatory difficulty of each sound is certainly a contributing factor (Kent 1992). Stop consonants require primarily a ballistic motion of one or two articulators which completely closes the vocal tract, but fricatives require precise control which only partially opens or closes the vocal tract. Shaping of the tongue is much more involved for liquids than for glides.

The above data suggests that anatomical constraints, articulatory complexity, and motor skill play a central role in explaining phonological development (e.g., Kent 1992, Oller & MacNeilage 1983).

2.2 Generativity and context-sensitivity

As children master the production of new subsyllabic sounds, they quickly generalize them to new lexical contexts. For example, Menn's (1976) subject, Jacob, pronounced /b/ as [d] in *byebye* [d&d&, dEda, dedo ...], *bang* [d&jN:], *beads* [di], *box* [da], and *ball* [dO] through 18 months. Starting with *bottle* [bAp?m] and *baby* [bajp?i], new /b/-words added to his vocabulary no longer showed this error (e.g.,

Table 2.1 Types of pronunciation errors

Class of Error	Type of Error	Examples
Substitutions	Fricatives → Stops	*soon* → [dun] *fly* → [baj]
	Liquids → Glides	/r/ → [j] /l/ → [w]
	Place-of-articulation errors	*ball* → [dO]
	Voicing errors	cat → [g&t]
Deletions	Simple deletions	*fish* → [IS] *shoes* → [uz] *soup* → [up]
	Cluster Reductions	*cheese* → [iz] *string* → [dIN] *stone* → [don]
Context Sensitive substitutions	Consonant harmony	*duck* → [gAk] *boot* → [bup]
	Metathesis	*snow* → [nos]
	Template matching	*telefóno* → [fa'tino]
Combinations of errors	Cluster reduction, consonant harmony, initial voicing	*stuck* → [gAk]

bump, *bus*, *block*, *brush*). Furthermore, most older /b/-words — those which had been pronounced with onset [d] — were corrected, with one notable exception.

Ball resists the change. This is especially puzzling because the first instance of *ball* [bo] on 17;13 was fairly accurate. But it did not reappear in Jacob's corpus until two months later at 19;13, and then only during a period of intense experimentation with *door* [do, dO, du, Adu]. Between 19;13 and 19;29, *ball* was pronounced [da, dO, dA, djA, djo]. It appears that *ball* was assimilated to *door*'s phonetic characteristics, not the phonetic features of the /b/-words. *Ball* shares a middle-back-rounded vowel with *door* which it does not share with the /b/-words. This is but one example of a word which does not generalize in ways typical of adult phonology. Generalization patterns are *context sensitive,* involving the whole syllable, not merely the consonant segment.

These data suggest a primitive *compositional structure,* one which does not correspond exactly to adult phonotactic structure. Children's speech is compositional because utterances are not memorized idiomatic forms; they are productively composed of component sounds according to specific phonotactic constraints, even if not in accord with adult phonotactic rules. The context sensitivity accounts for the ways in which the child's phonology differs from the adult's. The next several sections explore these ideas in greater detail

2.3 Relationships between adult target words and children's speech

Children's context sensitive substitution and deletion of segments which appear in adult target words is routine. However, it is the relationship between adult target words and children's realization of them and the regularity of this relationship which are special. Smith (1973) and Menn (1971) capture these effects with a set of *realization rules,* a formal device resembling the rewrite rules of adult phonology (Chomsky & Halle 1968). Realization rules relate distinctive features occurring in the adult target word

with those occurring in the child's realization and show the effect of context. Sometimes they are optional, thus do not express a perfectly lawful relationship between target and actual utterance.

In the following examples, Smith's (1973) subject, Amahl, transforms word-initial /s/ into different sounds in different contexts. The narrative attached to each example attempts to capture this in terms of rules operating in a fixed order of precedence. Several examples involve the process of consonant harmony, one of the more dramatic cases of the child's context dependency. Consonant harmony is the assimilation of place, manner, or other features of one consonant to match the same property of another, non-contiguous consonant in the same word.

- *stop* → [bOp, dOp] (stage 1)

  The /s/ is deleted. The remaining /t/ optionally harmonizes with the final labial consonant's place of articulation, but is voiced because it appears in onset position.

- *sock* → [gOk] (stage 1)

  The /s/ in *sock* assimilates the velar feature of the final /k/. The resulting velar fricative is transformed into a stop consonant, thence into a voiced onset consonant.

- *soap* → [u:p] (stage 1); [do:p] (stage 10)

  Early in Amahl's development, the initial /s/ is deleted by a rule which deletes the onset consonant in /sVC/ or /SVC/ syllables when it is not otherwise affected by velar or labial consonant harmony. Later in development, an initial fricative is transformed into the corresponding voiced stop consonant, i.e., /s/ → /d/.

- *sit* → [lit], *side* → [lait] (stage 7)

  When /s/ is followed by another alveolar consonant, it is transformed into a lateral /l/.

- *sap* → [w&p], *sip* → [wip] (stage 4); *some* → [wAm] (stage 9)

  The /s/ in these forms is first transformed to /f/ by labial harmony, then to [w] by a rule which transforms a word-initial labio-dental fricative to a sonorant.

2.3.1 Implications of realization rules

Realization rules and especially their successful application by Smith (1973) and Menn (1971) to describe children's early speech have several implications which constrain a model of phonological development.

- The approximately lawful relationship between child and parental utterances implies that some underlying representation of parental speech governs child speech.

- Children's pronunciation errors are sufficiently regular to be captured using such a formal device. This suggests either a rule-based or associative mechanism linking an internal representation of target words with their pronunciation.

- It also implies the compositionality of child's utterances. Adult words are composed from elemental sounds; children's utterances are also composed from elemental sounds. Thus, realization rules demonstrate that children's speech is not idiomatic, but that it is productively related to the phonetic patterns of target adult words. Compositional structure has further implications which are explored throughout this thesis.

- The rules' context dependency implies that the child's compositional structure does not match the parent's phonology. Specific patterns of context sensitivity suggest that the underlying representation of parental speech is suprasegmental and captures contextual features of the target sounds.

2.4 Syllabic and whole-word constraints

The correspondence between target and actual utterance is so indirect in some errors that it resists a segmental analysis. The order of segments or syllables may even be reversed as in the following *metathesis* examples:

snow → [nos]
'gator (alligator) → [d&ge]
cup → [bAk]
telefóno → [fa'tino]

Some such forms may best be described as conforming to a template, canonical form (Ingram 1974, Menn 1976), prosodic structure (Waterson 1971), or output constraint (Menn 1983; see also Ferguson & Farwell 1975, Macken 1979). That is, instead of the output being shaped to conform to the target word, phonetic features in the target word are shaped to fit a relatively stable output form. Nonetheless, some relationship between target and actual utterance is usually evident. Waterson's (1971) analysis is instructive in this regard. Table 2.2 summarizes correspondences between adult and child forms which fall into Waterson's "labial prosodic structure".

Consonant harmony is a relatively dramatic context sensitive error for which there is virtually no precedent in adult speech. It is the change of place, manner, and/or other properties of one consonant to match the same property of another, non-contiguous consonant in the same word. Harmony may involve repetition of a single feature (e.g., place or manner), the entire segment, or even affricates and clusters (Vihman 1978).

In the following examples of *progressive harmony*, the properties of the onset sound affect subsequent consonants.

boot → [bup]
snap → [n&t] (Menn 1971)

In *regressive harmony*, properties of later consonants affect earlier consonants. E.g.:

duck → [gAk]
tub → [bAb]
stone → [non] (Menn 1971)
tiger → [gaig6]
driving → [waibin] (Smith 1973)

Harmony's presence is betrayed by its systematic distribution, not by individual words. For example, Daniel's [gAk] is not simple consonant substitution because he pronounces open /dV/ syllables accu-

Table 2.2 Labial prosodic structure

Targets and outputs

Word	**Target**	**Child's Realization**
fly	flaI	1;5: w&/bB&, 1;6: B&/v&/bB&
barrow	'b&r^{w}6^{w}u	1;5: w&w&, 1;6: bAwU
flower	fla:/'flaw6	1;6: v&/v&w&

Features of adult targets and child's realizations

Adult Target Features	**Child Realization Features**
labiality at least at onset [f, b, r^{w}, w]	labiality at onset of each syllable [w, b, B, v]
continuance of most consonants [fl, r^{w}, w]	continuance of most consonants [w, bB, v, B]
openness of most vowels [a, a:, aI, &, 6]	openness of most vowels
prominence of 1st syllable	prominence of one syllable
syllable structure C(C)V(CV)	syllable structure CV(CV)
frontness of 1st syllable [fla, flaI, b&]	frontness of all syllables except [bAwU]
sonorant endings of all syllables	—
liquid feature [l, r^{w}]	—
centrality in 1 or 2 syllables [fla..., w6, r^{w}6^{w}]	—
non-rounding of 1st syllable [fl, b]	—

rately (e.g., *tea* [di]) but other velar-final forms (e.g., *book* [gUk]) with regressive harmony regardless of the initial consonant (Menn 1971).

Such data, either metathesis or consonant harmony, as well as the prosodic structure analysis offered by Waterson, suggest once again that the underlying representation of adult speech is at least supra-segmental and more likely includes the whole word.

2.5 Overgeneralization and inertia

Often errors are introduced when a newly mastered sound is generalized to new lexical contexts in ways which do not correspond to adult phonetic categories. For example, Jacob (Menn 1976) correctly used a palatal-/s/ combination [js] at the ends of several new words (e.g., *cheese, ice, noise, nice, toys*) but then overgeneralized [js] inappropriately in *bus* [bajs], *grass* [rajs], *mess* [majs], and *horse* [hOjs] (Menn & Matthei 1992). Daniel's nasal harmony started innocently enough with the nearly correct 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).

Nonetheless, the oldest, longest established forms often resist generalization by new sound patterns. Daniel's nasal harmony became more and more inclusive until even established forms for *down* [d&wn] and *stone* [don] were infected (e.g., [n&wn], [non]), but only after all the other words to which the rule could apply. The oldest forms show *inertia* in the face of generalization by new articulatory patterns.

Overgeneralization and inertia suggest the use of an associative model rather than a lookup table or a fixed set of rules. While associative models have the desired generalization properties, they are also subject to overtraining in which they memorize patterns subsequently resistant to change. The data further

reinforce the need for an underlying representation of parental speech which will induce the desired overgeneralization, one which has suprasegmental and contextual features.

2.6 Irregularities

Despite the overall regularity of the relationship between child and adult pronunciations, there are nagging irregularities. One example, a case of phonetic "crosstalk" between /Ej/ and /i/ drawn from Menn (1976) data, is presented in Table 2.3. These two sounds are often interchanged (as shown in bold). Target words appear in the top row, actual forms by age (month, day) and target appear below.

This is not a "classic" case of overgeneralization. Two articulatory patterns affect each other, and neither is stable. Irregularities like these (e.g., Menn & Matthei 1992, Priestly 1977) reinforce the case for an associative model. Moreover, they suggest a distributed or subsymbolic underlying representation of target words which captures the phonetic similarity of sounds. In this example, the diphthong [Ej] ends with a palatal which is nearly identical to the high front vowel [i].

Table 2.3 Crosstalk among several of Jacob's related words

Age	*tape*	*tea*	*key*	*cake*	*away*	*gate*
16;16		ti				
17;16	Ej, t?Ej					
17;18	t?Ej, dE	di, **dEj**	k^hi		AGE	
17;23	te		**xiE,** k^hi			
17;25	**?txi**		**xE**			
17;27	**tIjA, ti, di**	ti				
18;2	gEj					
18;18	**ti,** dZE		ki, xi		awE	**gi**
18;20	taj, dEj	ti, di	ki	**ki**, ke	we	
18;27	**ti**		ki			

2.7 Variability

Few of these pronunciation patterns are entirely stable. Realization rules are often optional. Utterances which conform to syllabic or whole-word constraints do so only approximately. Crosstalk and other irregularities are revealed in part by the competition among different realizations of each sound. Thus, *tape* varies between "tay" and "tee".

Granularity of phonetic variation in children's speech may be fine (perhaps evident only to the trained ear or acoustic instrumentation) or coarse (adult-like segments and context sensitive patterns). The following are fine-grained phonetic variations of 7 different intended consonants or consonant clusters in Jacob's speech.

/r/ → [r, r_0, l, w, j]
/l/ → [l, L, dZ, z, Z]
/S/ → [s, S, x]
/tr/ → [tS, tl_0, tw, k, g]
/k/ → [k, x, k^h]
/t/ → [t, d, t?, dZ, g]
/w/ → [G, w, B] (Menn 1976)

Coarser-grained variation often emerges when error patterns change. When old errors give way to newer errors or more accurate, adult-like pronunciations, old and new patterns often compete. Table 2.4 shows the alternation of Daniel's regressive velar harmony with consonant contrast as it gave way to the latter (Menn, unpublished data). Note the competition between the inaccurate /g/ onset and the correct forms which have [b], [p], [m], [d], and [t] onsets in words which end with [g], [k], or [N]. Harmonic forms are bold; correct contrasting forms are italicized.

Table 2.4 Variation in Daniel's regressive harmony as it crumbles

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"
33;4	bIwd A **gIg** t&wr	"build a big tower"
	Is nat tu *bIg*	"this not too big?" (about spoonful of something)

As apparent in Table 2.3, both fine-grained and coarse-grained variation may occur at the same time. For example, while the vowel and diphthong compete in Jacob's pronunciation of *key*, the articulation of /k/ also varies [k, x, k^h]. Only the syllabic structure remains the same. Likewise, pronunciation of /t/ in *tape* and the voicing of /t/ in *tea* are unstable.

Is the source of this variation articulatory, phonetic, or phonological? Surely consonantal variation often seems articulatory, revealing variations in timing, voicing, aspiration, place-of-articulation, magnitude of the vocal tract constriction, or some combination thereof. However, Daniel's regressive harmony cannot be articulatory. Word-onset non-velar stops [b, p, m, d, t, n] are mastered in other lexical contexts. Thus, data on variation suggest not only that output is stochastic, but also that variation may have two sources — articulatory and compositional.

2.8 Perception of speech and underlying representations

The foregoing analysis suggests that some representation of adult speech underlies children's speech (Section 2.3.1). What form does it take? Modern generative phonology since *The Sound Pattern of English* (Chomsky & Halle 1968, see also Durand 1990) has assumed that superficial adult speech patterns are governed by underlying abstract representations and their contextual interactions (e.g., nasal assimila-

tion, vowel harmony). However, children's speech seems to be governed by a different set of underlying forms and by different rules (see also Section 2.4 and Section 2.5).

2.8.1 Representation of target utterances

A working hypothesis which guides most phonological research is that by the time children start talking, they have a more-or-less accurate phonetic or phonemic representation of adult speech. This is based on the observation that passive vocabulary (recognition of words) precedes active vocabulary (production of words) statistically (Benedict 1979) and individually in longitudinal studies (e.g., Menn 1976). In considering the possible sources of his son's speech errors, Smith (1973) argues that perceptual errors are unlikely because Amahl responds appropriately to conversation and verbal instructions. Smith also informally tests his son's discrimination of minimal pairs of mispronounced words (e.g., *mouth* vs. *mouse*), a procedure applied more thoroughly by other investigators (Section 2.8.2).

Consistent errors in pronunciation and the close relationship between child and adult sound categories revealed by realization rules imply accurate phonetic encoding of adult distinctions. One of Smith's (1973) examples is instructive. Amahl transforms coronal stop consonants /d,t,n/ which precede laterals into velar stop consonants /g,k,N/. Thus, *pedal* becomes [bEgu] and *bottle* becomes [bOkl]. Delayed-release or fricative coronals /s,z,tS,dZ/ which do not occur at the end of a word are transformed into corresponding stops. Thus, /s/ becomes /t/ in *whistle* [witl] and is deleted from *pistol* [pitl]. Paradoxically, Amahl pronounced *puddle* as *puggle* [pAgl], and *puzzle* as *puddle* [pAdl]. Their distinct pronunciation implies that the sounds are distinguished perceptually. The example also suggests that the key to understanding children's speech errors is not just perceptual and not just articulatory — Amahl's problem is not in articulating *puddle* — but the relationship between listening and speaking, including how contextual information is encoded and expressed.

2.8.2 Phonemic vs. phonetic perception

Phonemic distinctions are the dimensions of sound which encode meaning. Phonetic distinctions denote differences in articulation of sounds. Several studies ask if children's representations are phonemic and whether such phonemic representation is a prerequisite of speech production (Barton 1976, Edwards 1974, Shvachkin 1948/73). This is a stricter test of perception than whether children can distinguish sounds phonetically. It asks whether children store sounds according to the minimal distinctions between contrasting pairs of words such as *mouth* vs. *mouse* (dental vs. alveolar), *pat* vs. *bat* (voicing), *goat* vs. *boat* (velar vs. labial).

Children's vocabularies do not generally contain many contrasting pairs, thus these studies attempt to teach children to recognize new contrasting words and nonsense words. Despite experimental design problems and subsequent methodological questions (Barton 1980, Ingram 1989), results generally suggest that "phonemic" perception of a given sound difference precedes the correct production of the difference.

Do these studies measure true phonemic distinctions or only phonetic distinctions? Phonemic categories are more inclusive than phonetic categories. The *phoneme* /t/ includes several *allophones* (categories which differ phonetically but do not signify semantic distinctions): the plosive, voiceless, aspirated word-initial stop consonant [t^h] in *tick*, the voiceless but unaspirated stop [t] which follows /s/ in *stick*, the glottal stop which often accompanies word-final closure in *hit* [hi?^{t}], or an apical flap as in *butter.* The Shvachkin, Edwards, and Barton studies do not detect whether children group allophones, thus test necessary but not sufficient conditions for phonemic perception.

More direct studies of infant perception imply that allophones are grouped together. More precisely, native and nonnative consonants are assimilated to and perceived as instances of sound categories

which are phonemic in the child's language (Best et al. 1988, Werker & Pegg 1992); vowels are also assimilated to the nearest phonemic category (Grieser & Kuhl 1989, Kuhl et al. 1992). Soon after birth, mere acoustic contrasts are easily distinguished, but before the end of the first year, there is a declining ability to distinguish sounds in the same phonemic class, (although with training adults can learn to distinguish allophones; Werker & Pegg 1992).

The question still persists whether the sort of perception identified in these studies is truly "phonemic." A true phonological system is built only gradually (Nittrouer et al. 1989). An infant of 10 months, whose passive vocabulary is still rather tiny and whose active vocabulary is at best only a few words, is unlikely to have more than a rudimentary system (Werker & Pegg 1992).

Werker and Pegg suggest and review several alternative hypotheses which might overcome this quandary, including perceptual tuning, articulatory mediation, cognitive recategorization, and self-organized learning. They argue that a successful account must explain (1) the timing of selective categorical perception (6 months for vowels, 12 months for consonant-vowel syllables), (2) the conformity of reorganized phonetic categories to phonemic categories, and (3) the ability of adults to switch between phonemic and phonetic listening under the right conditions.

Part of the problem is confusion about what "phonemic" means in this context. The term implies (1) the size of a segment of sound, (2) categories of sound which contrast linguistically, and (3) the minimal linguistic unit which encodes semantic information. For an adult, these three characteristics intersect, but that is not necessarily the case for the child. Perceptual experiments cited above address only perceptual assimilation and contrast, not segment size or semantic encoding.

Self-organizing processes that learn to model the statistical distribution of speech sounds in the linguistic environment, and particularly those which model sounds as if they were generated by a mixture of stochastic processes (e.g., Nowlan 1991a,b), may account for such phoneme-like categorization of sounds. With such models it may be possible to account for both the phonetic variation (and assimilation) which occurs within each "phonemic" category and the phonetic contrast which occurs among "phonemic" categories. This is based on the assumption that phonemically distinct sounds occur as distinct statistical distributions in the linguistic environment, but acoustically distinct sounds which are part of the same phoneme will fall inside the same statistical distribution (holding syllabic context constant). In natural linguistic settings, different allophones occur in different syllabic and morphemic contexts. Children may classify allophones distinctively, especially if phonetic perception is based in part on dynamic spectral features (see Section 2.8.3). Although they are grouped together as the same perceptual entity by adults, it is implausible that children group them together without additional knowledge of semantic encoding. Consonant categorization experiments do not test this aspect of phonemic categorization. Instead, they test the perception of non-native allophones or the introduction of allophonic segments into unnatural syllabic contexts. Furthermore, there is no evidence that a child must recognize allophone variations as the same perceptual entity in order to learn how to pronounce them.

2.8.3 Syllabic and dynamic phonetic perception

Analysis of children's speech (see Section 2.4 and Section 2.5) suggests that underlying representations are at least suprasegmental and possibly syllabic or morphemic. It is not clear from speech production data alone whether this suprasegmental organization is perceptual, motor, or both.

It has long been conjectured that the syllable, not the phoneme, is the critical segment of *adult* speech perception (e.g., Neisser 1967). When speech is presented monaurally but switched between the ears, it is least intelligible if interrupted at least once per syllable (Huggins 1964). Adults are able to identify place-of-articulation even for stimuli which lack invariant static spectral properties (Cooper et al.

1952, Delattre et al. 1955), thus seem to depend on dynamic syllabic cues. Indeed, attempts to find invariant static spectral cues for stop-consonant identification have failed (Blumstein et al. 1982). Rather, dynamic spectral features seem essential for consonant or vowel identification (Furui 1986, Lahiri et al. 1984, Lindblom & Studdert-Kennedy 1967, Kewley-Port et al. 1983, Nossair & Zahorian 1991, Strange et al. 1983, Walley & Carrell 1983). Particularly salient are the spectral transition between consonant and vowel and the dynamic spectral features during and immediately following a stop-consonant's burst.

Recent evidence increasingly suggests that children's phonetic perception is organized around the syllable and its dynamic properties, even at a very early age. In a habituation experiment involving a list of four syllables, two-month old infants notice the insertion of a new syllable into the list even if it differs from the others in a single consonant or vowel (Bertoncini et al. 1988, Jusczyk & Derrah 1987). Even four-day old infants notice the introduction of a syllable which differs from the habituated syllable by at least a vowel. In habituation experiments involving two-syllable utterances, newly introduced bisyllables which included an already familiar syllable aroused less interest than bisyllables composed of novel syllables, *even if the novel syllables shared consonants and vowels with familiar syllables* (Jusczyk et al. 1995). Infants apparently perceive and distinguish syllables holistically, without regard to segments they may have in common.

Infants are able to discriminate very short CV syllable onsets spliced from full syllables (Bertoncini et al. 1987). Infants able to discriminate [pat] and [tap] show considerable difficulty in distinguishing [pst] and [tsp] unless a vocalic context is added to form [upstu] and [utspu] and thus introduce syllabic structure (Bertoncini & Mehler 1981). Young children focus on spectral changes rather than static acoustic information (Walley & Carrell 1983) until rather late in development (Nittrouer 1992). Indeed, spectral transitions appear to be more salient for children than for adults (Walley & Carrell 1983).

2.8.4 Model constraints based on speech perception data

In summary, a number of conclusions may be drawn from children's speech perception which suggest further constraints on a model of phonological development.

- Data on adult and infant speech perception confirm the syllabic or suprasegmental representation of parental speech which is revealed by children's speech production.
- Speech perception is contextual; dynamic spectral features are particularly salient.
- Phonetic categories are learned early, suggesting a self-organized learning process.
- Categories approximately correspond to phonemic contrasts in the linguistic environment, suggesting a statistical mixture model which accounts for variation and assimilation of sounds within each category and the distinctiveness of sounds across categories.
- Perception is nearly veridical, a faithful though compact, categorical representation of sounds in the linguistic environment.

Chapter 3 Computational Models of Phonological and Articulatory Development

This chapter surveys relevant models of speech perception and production, focusing on phonological, articulatory, and related motor control models, their possible contribution to a more complete theory of phonological development, and what problems remain unaddressed.

3.1 Speech perception models

The apparent syllabic organization of speech perception leads Mehler et al. (1990) to propose a word recognition model in which a bank of syllable recognizers activate lexical and phonological patterns. Jusczyk's (1993) WRAPSA is a more complete conceptual model of word recognition. Acoustic features sufficient to explain the early discriminability of most sounds are extracted from the speech signal in the first processing stage. To explain learned categorization of speech sounds and assimilation of unfamiliar sounds to such categories, the second stage weights features according to their prominence in the linguistic environment. A third stage attempts a segmentation into word-sized or syllable-sized units which act as probes into an exemplar-based lexicon.

WRAPSA is not implemented. Several of its details remain somewhat vague, but it provides a framework which helps explain developmental phenomena. How to fill in some of the missing details is suggested by Sawusch's (1986) process model of adult speech perception. It proposes auditory feature detectors which might be incorporated into a developmental model, including dynamic and contextual features.

With his model, Sawusch attempts to resolve the problem of invariant perceptual cues — why the same static acoustic cue may give rise to different percepts in different contexts, and why different cues may give rise to the same percept in different contexts. He argues that segmenting speech into static, discrete units is what makes it so difficult to find invariant acoustic cues (Blumstein et al. 1982). Citing psychophysical experiments, he proposes that perceptual constancy results from the interaction of distinct auditory and phonetic coding processes. His model has four basic components: (1) spectral processing (with critical band filters and other signal processing components), (2) local auditory coding of static spectral and source (voicing, friction) features, (3) contextual coding of dynamic spectral patterns, and (4) phonetic integration, in which local and contextual features are paired with phoneme labels by a generate-and-test procedure.

Sawusch's model is not developmental. Jusczyk's model does not specify an algorithm by which acoustic features might be weighted and phonetic categories formed. To be complete, a model of phonological development must address these issues. It must also determine how auditory perception and motor control might be integrated.

3.2 Two classes of speech production models

Two contrasting goals characterize extant models of articulatory and phonological development and related models of articulatory control. Given a set of atomic sound units (phonemes, distinctive features, or other discrete representations), a model of abstract phonological competence attempts to account for their composition in ways that conform to the language's regular phonotactic structure or in ways that predict errors observed among adults and children. Given a continuous stream of sound, a model of articulatory skill attempts to determine the sequence of articulatory motions necessary to reproduce it and the

patterns of coarticulation which result when two sounds are combined. We evaluate each class of models in the following sections.

To our knowledge, no model that attempts to do both has been implemented. Yet, a complete model of phonological development must bridge both. Clearly, children learn what vocal tract, glottal, and pulmonary motions to make in order to pronounce the sounds of their native language. Also, children learn how to compose smaller sounds into larger ones. Part of a child's competence is articulatory, part is compositional.

3.3 Abstract phonological competence models

Several abstract phonological competence models specifically address development. Most deal with issues of morphophonology. These include Rumelhart and McClelland's (1986) model of English past tense acquisition, the several connectionist refinements which appeared after Pinker and Prince's (1988) critique (Daugherty & Seidenberg 1992, MacWhinney & Leinbach 1991, Plunkett & Marchman 1991), and a recent symbolic competitor (Ling & Marinov 1993). In addition, there are models of German article declension (MacWhinney et al. 1989), English plural and non-assimilatory suffixation (Lee & Gasser 1992), and receptive morphology (Gasser 1993). Several additional models are not intended as developmental accounts but use connectionist methods to discover underlying abstract representations which explain strictly adult data (e.g., Gasser 1992, Gasser & Lee 1990, Hare 1990, Touretzky & Wheeler 1990, 1991).

A number of mechanisms are exploited to mimic real phenomena. The morphophonological models attempt to elicit rule-like and exceptional behavior in the composition of tensed verbs, plurals, and articles. Connectionist models do so by balancing a network's generalization and memorization properties. Feature assimilation in vowel harmony is modeled by Hare (1990) using smoothness constraints and underspecified training signals in a recurrent neural network (Jordan 1986). Feature assimilation is involved in affixation as well (e.g., voicing in the regular English past tense and vowel quality in the irregular English past tense). The past tense models are trained with the correct pattern, but some past tense model feature representations are designed to encourage assimilation. Gasser's (1992) abstract representation discovery procedure elegantly exploits a recurrent network's discovery and distributed representation of temporal patterns and the network's pattern completion capabilities.

Key to the successful implementation of compositional structure in these network models are three factors. (1) The phoneme's phonetic properties and contextual relationships are represented by a distributed feature vector. (2) Compositional structure and phonotactic constraints are implicit either in network architecture or in contextual feature vectors. (3) An encoding relates discrete phonemic units with the network's internal representation. Rumelhart and McClelland use Wickelfeatures (after Wickelgren 1969) to represent the temporal order of phonemes. In MacWhinney and Leinbach's model, input and output units represent relative syllable position of phonemes. Gasser (1992) and Hare (1990) use recurrent network designs which discover and use a distributed representation of the temporal context.

Relevant to the present work is the ability of these abstract phonological models to learn compositional structure and exhibit rule-like behavior. Key mechanisms are the generalization and pattern completion properties of neural networks.

Also relevant is built-in knowledge of (1) phonemic segmentation, (2) phonotactic constraints (except for some recurrent networks models), and (3) phonetic and articulatory properties of phonemes. Knowledge of phonotactic constraints and temporal structure is learned only in the case of recurrent network models. For models of morphophonological and other relatively late developmental phenomena, it is appropriate to assume this as prior knowledge.

A model of the articulatory and phonetic foundations of phonology does not have this luxury. It cannot assume phonemic segmentation, because the context sensitivity of children's pronunciation errors and properties of children's speech perception suggest that children have a syllabic, contextual representation of speech (see Chapter 2). Such a model cannot use prior knowledge of phonotactic and phonetic properties of phonemes. Instead, it must explain how they are learned. Likewise, such a developmental model should presumably explain how correlations between auditory and articulatory events are learned. Prior knowledge of such correlations is implausible, in part because articulatory patterns are unmastered and phonetic categories are unlearned, and in part because the motor equivalence of many sounds makes a well-defined relationship between articulatory and phonetic patterns impossible (see also Lindblom 1991). Distinctive features (Chomsky & Halle 1968), traditional articulatory features (Ladefoged, 1972), or any of myriad alternatives (e.g., Shillcock et al., 1992), which are framed largely in articulatory terms, assume prior knowledge of articulatory-auditory correlations and are inappropriate as a representational basis for a model of the phonetic foundations of phonology.

A model of phonological development must bootstrap itself upon exclusively acoustic cues and their statistical distribution in the environment, aided only by architectural, representational, and algorithmic constraints, plus knowledge of which broad types of sounds are worthy of attention. Lindblom (1992) offers evidence that self-organization of phonological units is feasible, based on constraints which minimize articulatory costs, maximize discriminability of acoustic properties, and maximize generalization of gestural parameters among phonological units.

3.4 Articulatory skill models

Articulatory skill models address one of the most basic problems of speech production — what sequence of articulatory motor commands is required to generate a continuous stream of speech. This is often split into kinematic and dynamical subproblems. The kinematics of speech addresses the mapping between articulatory and acoustic coordinates. The speaker's *proximal* task is to generate a trajectory of vocal tract and respiratory motions. The speaker's *distal* task is to generate a sequence of sounds. That is, the goal and the success or failure of the immediate articulatory task is not measured in proximal articulatory coordinates but with respect to a different, acoustic or phonetic coordinate system. The dynamics of speech addresses the mapping between a sequence of possibly discrete motor commands or articulatory gestures and the more-or-less continuous articulatory or acoustic reference trajectory which defines the proximal or distal task.

3.4.1 Kinematics of articulatory and acoustic trajectories

For the naive agent (e.g., a child learning to speak), the distal task (the target sound to be produced) is known, but the proximal task (the articulatory motions which generate the sound) and the mapping between the proximal and distal tasks is unknown. However, the mapping from articulatory to acoustic space may be experimentally determined. That is, a *forward model* may be built which for each articulatory (and vocal tract) configuration identifies the corresponding acoustic sensation. An inverse of this mapping is required for kinematic control. An *inverse model* maps target acoustic sensations to vocal tract configurations. Unfortunately, the mapping from articulatory configurations to acoustic sensations is many-to-one (Atal et al. 1978). For example, the lengthening of the vocal tract to generate the sound [u] may be accomplished either by protruding the lips or by lowering the glottis (Ladefoged 1982). This is also called the "motor equivalence" problem. No unique inverse of the forward model exists, and an inverse model may not be determined directly.

Jordan and Rumelhart (1992) propose a method by which an inverse model may be learned even if the environment is characterized by a many-to-one mapping between actions and sensations. This indirect approach is applicable to speech and more generally to any similarly ill-defined kinematic or dynamic control problem. In this domain, the relationship between articulatory configurations and acoustic consequences is ill-defined, but the relationship between *articulatory errors* and *acoustic errors* is well-defined. The idea is to incrementally learn a particular inverse by exploiting this relationship. To learn the inverse model, articulatory errors (which are not known) must be corrected. This is accomplished by using the forward model to transform acoustic errors (which are known) into articulatory errors, which are then used to correct the inverse model. First, a neural network learns the system's forward model, i.e., learns to predict the acoustic consequences of each articulatory configuration. Once the forward model is learned, it may be used to convert acoustic errors into articulatory errors and indirectly train an inverse model. The inverse model is implemented with a second neural network whose inputs represent the target acoustic state and whose outputs represent the articulatory configuration.

Coupled, the two models behave as follows. Given a target acoustic state, the untrained inverse model computes its best guess of the corresponding articulatory configuration. The acoustic error is the difference between the target acoustic state and the sound emitted by the inverse model's chosen articulatory configuration. The inverse model, whose outputs are in articulatory coordinates, cannot use this distal error signal (i.e., an error in a different set of coordinates). However, the acoustic error is transformed into articulatory coordinates by back propagating it (Rumelhart et al. 1986) through the forward model. The articulatory error thus obtained is used to train the inverse model, incrementally approximating a particular inverse.

3.4.2 Forward model error inversion and articulatory skill

Among models of articulatory skill are Hirayama et al.'s (1992, 1993, 1994) physiologically based speech synthesis model and Laboissiere's model of vocal tract control (Bailly et al. 1992, Laboissiere et al. 1990, Laboissiere 1992). Each employs an acoustics forward model.

Laboissiere's model is most relevant to this thesis. It learns jaw, tongue, and lip motions necessary to pronounce a sequence of vowels. The distal teaching signal is a partial reference 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. Given additional smoothness constraints (Jordan 1990) and the manner in which the reference trajectory and errors are determined, the model learns to interpolate smoothly in the intervals between target vowels. Once trained, the model simulates skilled human behavior in classic bite-block and coarticulatory experiments.

3.4.3 Dynamics of articulatory skill acquisition

A kinematic inverse model is not sufficient to solve the problem of articulatory skill acquisition. Dynamical properties of speech articulators (damping and stiffness of jaw, tongue, lip, and other articulators) ensure that the translation from a sequence of motor commands to the articulatory or acoustic reference trajectory which defines a proximal or distal task is not direct. The acoustic consequences of each motor command are delayed and distributed over a period of time, overlapping with the acoustic consequences of other motor commands. Control of speech requires that responsibilities for acoustic patterns which overlap in time be assigned to individual articulatory actions. This is the *temporal credit assignment problem* (Sutton 1984). Speech production and perception involve several additional factors which further complicate temporal credit assignment. For example, acoustic feedback is delayed relative to motor commands because of the time needed to propagate and realize motor commands and analyze auditory feed-

back. The acoustic reference signal may be discontinuous. It may be completely missing, replaced by a compact, categorical encoding of the utterance.

The acoustic trajectory's discontinuity is especially acute for voiceless stop consonants, which are inaudible except after the stop's release. Acoustical information during a voiced or nasal stop is nearly as impoverished. One may infer a stop consonant's place of articulation from context, but only after a delay. For example, to pronounce the opening sounds in *kick*, the tongue body first moves up to contact the velum. Then, with lungs relaxed and velum closed to block passage of air through the nasal cavity, the tongue body is lowered. A sudden rush of turbulent air flows through the open glottis, then the glottis is closed until the vocal folds start to vibrate. Acoustic cues for identifying the initial consonant's place-of-articulation are the dynamic properties of the first 20-40 msec of aspiration (Kewley-Port et al. 1983). Actions to fill and relax the lungs, open the glottis, close the velum, raise the tongue body to the velum, and lower the tongue body all occur before there is any acoustic feedback.

Laboissiere's (1992) model learns stop consonants by using the locus (Delattre et al. 1955) as a teaching signal. The locus is the set of formant frequencies which is obtained by extrapolating formant trajectories forward or backward in time to the hypothetical time of articulator closure. This method thus avoids the difficult dynamical control problem. It is justified only if an inference method exists for estimating the locus, not a plausible scenario for a naive infant.

Parental feedback is not always present. Furthermore, adult and probably children's auditory echoic memory is short-lived (Baddeley 1992). Might long-term memory traces of target utterances replace direct supervision? The requisite memory storage makes this implausible. Besides, the evidence (see Chapter 2) suggests that children's accessible representations of adult speech are compact categorical forms which capture certain linguistically invariant phonetic features. They lose the richness of the original stimulus and bear little resemblance to continuous acoustic reference signals of the sort presumed by customary control models. Replacing the acoustic reference signal with a compact encoding of the target utterance or classes of target utterances adds a new dimension to the temporal credit assignment problem as well as a new decoding problem.

3.4.4 Inverse dynamics with temporal credit assignment problems

The solution of inverse dynamics or temporal credit assignment problems may be found using either supervised or reinforcement learning. Supervised learning is typically applied to tasks in which there is an explicit teaching signal and a signed error vector. In dynamic system control problems, the teaching signal typically consists of a reference trajectory (in proximate or distal coordinates), or via points (a periodic sampling of points) along a reference trajectory. Reinforcement learning is typically applied to tasks in which there is a set of scalar evaluative signals. Solution of the problem in either paradigm involves optimizing an objective function which measures the cumulative error or payoff incurred during the task.

Supervised learning of dynamic control may be decomposed into forward and inverse models in the same manner as for kinematic control. The forward dynamic model learns the mapping from a sequence of motor commands to an articulatory trajectory. The inverse dynamic model learns the inverse mapping (once the forward model is mastered) by back propagation through time in a recurrent network which incorporates both forward and inverse models (Jordan 1990). The high computational cost of a recurrent network might be avoided by a feedforward dynamical model (e.g., Hirayama et al. 1994), but only at the expense of ignoring articulatory feedback.

With reinforcement learning methods, the controller learns an extended plan of action which maximizes the likelihood of future success as measured by cumulative discounted payoff. There is no distinction between forward and inverse models. However, a reinforcement-based controller does not learn a

direct inverse and does not stochastically search over all possible trajectories. Rather, it uses an asynchronous version of dynamic programming to incrementally build a model of the value of each state and the marginal value of each action in each state, caching knowledge won from partial trajectories and building thereon.

Jordan and Rumelhart (1992) distinguish between the problems and methods of supervised and reinforcement learning paradigms. While it is possible to use reinforcement learning methods to solve a supervised learning problem by translating signed error vectors into some corresponding evaluative signal, one loses the directional information contained in the error vector. It is also possible to use supervised learning methods to solve a reinforcement learning problem by using a forward model to translate evaluative signals into signed error vectors (Jordan & Jacobs 1990, Markey & Mozer 1992, Munro 1987) but at the cost of the additional forward model.

Jordan and Rumelhart point out that when a task combines dynamical properties and delayed consequences, building a forward or inverse dynamics model may be infeasible if the environment is too complex or wasteful if the details of the system's trajectory are not needed for its control. Instead, they suggest using an *integrated quantity* which measures overall task success. Surely, the task may be learned by using a forward model of the integrated measure (e.g., Jordan & Jacobs 1990), but only with the added expense of the forward model. This may be avoided with reinforcement learning methods which use the same integrated measure as an evaluative signal.

Similar circumstances apply to control of natural speech. There are two ways in which phonetic feedback may be used to learn the control of natural speech: (1) as an integrated quantity which encodes the success of the entire task, or (2) as an explicit but delayed teaching signal from which one might be able to derive a signed error vector. In the latter case, the controller must solve articulatory dynamics, inverse acoustics, the additional temporal credit assignment problems due to delayed auditory feedback, and must also decipher the compact phonetic representation in which the feedback is encoded (see Section 3.4.3 and Chapter 2). It is not clear whether reinforcement or supervised methods are more appropriate. In Section 3.5, we review reinforcement learning methods in greater detail, and we return to this question again in Chapter 7.

3.4.5 Dynamic components of articulatory skill models

Models of articulatory skill for the most part only minimally incorporate dynamical components. Laboissiere et al. (1990) implements a model which is purely kinematic save its smoothness constraints and extrinsic timing signals.

Hirayama et al. (1992, 1993) implement a forward dynamics model of the muscular-skeletal system which learns the mapping between electromyographic (EMG) activity and articulatory trajectories. Recently implemented is also a feedforward inverse model of muscular-skeletal dynamics that generates EMG signals and a smooth articulatory trajectory matching a sequence of discrete articulatory target configurations (Hirayama et al. 1994). Articulatory targets are not determined from an integrated model of forward acoustics but are chosen independently. Furthermore, the model is strictly feedforward to avoid sensory feedback delays. This seems justified in a model limited to muscular-skeletal dynamics. However, it cannot respond to environmental perturbations known to affect natural speech (Abbs & Connor 1991, Kelso et al. 1984).

The teaching signals for each of these models are reference trajectories — articulatory via points in Hirayama (1994) and partial acoustic trajectories in Laboissiere et al. (1990). Temporal credit assignment is thus considerably simpler than if the teaching signal were based on delayed phonetic feedback.

3.5 Reinforcement learning as an alternative control strategy

An alternative approach for the control of speech relies on reinforcement learning, in which the controller learns an extended plan of action which maximizes the likelihood of future success.

In a reinforcement learning paradigm, a controller receives only a scalar reinforcement signal from the environment evaluating its actions, rather than a teaching signal specifying its precise error. The reinforcement signal is often delayed, requiring the controller to determine responsibility for its actions not just structurally but also temporally (Sutton 1984). Algorithms for learning from delayed signals (Barto et al. 1990) are based on principles of dynamic programming (Bellman 1957, Bertsekas 1976). They learn to incrementally predict future rewards by temporal-difference learning methods (Sutton 1988). Instead of predicting the final outcome of a sequence of events, such methods learn only how the outcome's prediction changes from step to step. Working backwards from the goal, reinforcement learning methods determine the best sequence of actions to get to the goal.

One such method, *Q-learning* (Watkins 1989), 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. Q-learning has been shown to converge to an optimal plan for a finite Markov decision task under a number of specific conditions (Watkins & Dayan 1992). In a *Markov decision task* (see Figure 3.1), an agent seeks to control a finite-state, discrete time, stochastic dynamical system. At each time step, the agent observes the current *environmental state* **x** and executes an *action a*. The system thence makes a probabilistic *transition* to state **y**, and the agent receives a *reinforcement r* which is a function of the previous state **x**, new state **y**, and action *a*. This transition is a function only of state **x** and action *a*, not of any previous state (summary due to Singh 1992d). The agent's goal is to determine a control policy that maximizes some *objective function*, typically a cumulative measure of reinforcement over time.

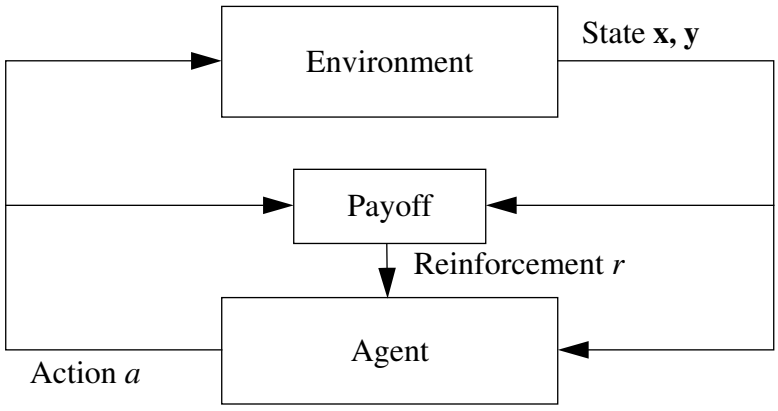


Figure 3.1 Components that define a Markov decision task

In Q-learning, the agent incrementally learns a function $Q(\mathbf{x}, a)$, which for every state **x** evaluates the expected utility of performing each possible action a_i. 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, $Q(\mathbf{x}, a)$ (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 thereafter.

3.5.1 Advantages and disadvantages of reinforcement learning

The chief advantage of reinforcement learning is its specialization for solving problems of temporal credit assignment. The principal disadvantage is that reinforcement learning (and specifically Q-learning) is untested in the control of speech and in most other large-scale domains. It is not thought to scale up well, and it is not proven to work with methods of function approximation (Watkins 1989). To extend its reach to problems which have large domains (such as speech), some form of generalizing function approximation is essential and has been tested in a number of domains (e.g., Lin 1992, 1993; see also Anderson 1987, Tesauro 1992). Under certain conditions, Q-learning provably fails with function approximation (Thrun & Schwartz 1994), but judicious choice of input representation may have considerable benefits (Moody & Tresp 1994).

Another disadvantage is the impracticality of using standard Q-learning in high-dimensional action spaces. For example, even the simplest articulatory model of the vocal tract involves at least a dozen dimensions of motor control involving jaw, lip, tongue body, tongue tip, velum, hyoid, glottal, and lung motions. Solutions are proposed by Markey (1994a, see also Chapter 7), Tham & Prager (1993), and Baird & Klopf (1993).

Despite the mathematical limitations, artificial neural networks and reinforcement learning methods have been used together with considerable success in game playing (e.g., Boyan 1992, Lin 1992, Schraudolph et al. 1994, Tesauro 1992), robotics (e.g., Anderson 1987, Gullapalli 1993), and animal learning (e.g., Doya & Sejnowski 1994, Montague et al. 1993, Sutton & Barto 1990).

3.5.2 Models of hierarchical reinforcement learning and control

One method of extending the reach and power of reinforcement learning systems is to decompose tasks hierarchically. Without such decomposition, some composite tasks may be intractable (e.g., Mahadevan & Connell 1991). Three principal approaches to hierarchical reinforcement learning have been reported. Singh's (1992a) CQ-L adaptation of Q-learning views a composite task as a single Markov decision task to be decomposed by subtask using a mixture of experts, each learning the environment of a subtask. Dayan and Hinton's (1993) feudal reinforcement learning distributes a task among a modular hierarchy of agents and subagents that have been pre-assigned to portions of the state space. A third approach is more heuristic, drawing on domain-specific knowledge to decompose a complex task (Kaelbling 1993, Lin 1993a,b, Mahadevan & Connell 1991). An examination of these approaches reveals six principal dimensions along which hierarchical reinforcement algorithms may vary. Choices along these dimensions are not entirely independent.

Division of labor. Labor may be divided by state or by task. When labor is divided according to state, each managing agent's role is to choose the action to be performed by the subagent preassigned to the current state, and each subagent learns to do a diversity of jobs in its assigned state. When labor is divided by task, each subagent learns to do one job well, and each managing agent's role is to choose the subagent most likely to behave optimally given the managing agent's own observed state.

Relationship between agent and subagent. The relationship among agents in one level and subagents which they manage may be supervisory or delegatory (Watkins 1989). The supervising agent retains complete initiative over its subagents. The delegating agent passes control to a subagent, whose responsibility it is to pass control back to its manager when its subtask has been completed. Delegatory control seems best suited to a division of labor by state. Otherwise a subagent has no principled method to decide when to relinquish control except to learn when a subtask is complete, which further complicates its job.

Abstract state discovery. State specialization requires division of the environment into abstract states, such as maze subregions in Dayan & Hinton's domain, application subspaces in Lin (1993a,b),

applicability conditions in Mahadevan & Connell (1991), or regional landmarks in Kaelbling (1993). Even task specialization requires some advance knowledge of which states represent subgoals (Singh 1992a). Decomposition into abstract states or abstract tasks may be adaptive or engineered in advance from pre-existing knowledge. Singh (1992b,c) proposes a method for discovering abstract states. However, he argues that a completely general method for decomposing problems hierarchically is unlikely to be tractable and bootstraps his system with elemental tasks.

Distribution of rewards among control levels. The distribution of rewards depend on the specialization method and the nature of the problem to be solved. In state specialization systems such as feudal learning, the subagent is rewarded when conforming to a managing agent's request, and the overall manager is rewarded when achieving the overall goal. For task specialization systems, the managing agent is rewarded based on the adaptiveness of its choice of a subagent, but reward of subagents is more complicated. Singh (1992a) assumes that no intermediate rewards are received as subtasks are completed. His system thus faces a difficult temporal credit assignment problem and for this reason views the composite task as a single Markov decision task.

Relationship of states among control levels. Reinforcement learning is guaranteed only for a Markov process and when several other conditions are met (Watkins & Dayan 1992). In order to conform to the Markovian property, the future sequence of states must be exclusively a function of an agent's current state, its own future actions, and the unchanging transition probabilities which characterize the environmental dynamics. If agent and subagent share the same state, the actions of one may compromise the effect of the other's actions, violating the constraint. In the feudal system, agents observe the environment only at the granularity which is affected by their own actions; they do not observe the irrelevant changes in the environment due to finer grained action by their subagents. The CQ-L system's experts (analogous to subagents) observe the physical state, but the gating module (analogous to a managing agent) observes only information about subtask completion. In each system, domain knowledge is used to engineer an artificial separation of states.

Application of domain-specific knowledge. Domain specific knowledge is required by each decomposition strategy — to determine division of labor, to choose abstract states or the dimensions along which states are abstracted by automatic means, to decompose tasks into subtasks, to design subtask-specific reinforcement functions, and to determine how to distribute rewards.

Chapter 4 HABLAR's Architecture and Behavior

Empirical findings about developmental phenomena constrain HABLAR (Chapter 2). Problems of speech production and perception not comprehensively addressed by existing cognitive models motivate it (Chapter 3). To start this chapter, we summarize the motivation and constraints. To address them, we introduce HABLAR's architecture and basic operation, illustrating the model's operation with how it might imitate a one-syllable word.

4.1 Empirical constraints and problems of speech production

Two basic problems of human speech production face any theory of phonological development. First, given a continuous stream of sound, what is the sequence of articulatory motions necessary to reproduce it? Second, how is speech structured into elemental units, and how are the units composed into unique utterances? A solution of the first question will account for the relative articulatory difficulty of different sounds revealed by children's speech. An answer to the second question will account for the way in which children recombine already mastered sounds into unique utterances and why their phonotactic patterns do not correspond to adult phonology.

The solutions, however, must be integrated into the same model. Although existing models of articulatory skill and abstract phonological competence address one question or the other, no model seems to address both. Moreover, no existing model apparently addresses the role of auditory perception in the control of speech, especially how it shapes the underlying representations which govern the production of speech. A complete model must address these questions and satisfy the empirical constraints identified in Chapter 2.

In particular, children's speech production and perception imply that some veridical underlying representation of speech governs children's imitation and pronunciation of target words. It is apparently a learned representation assimilates acoustically similar sounds into existing categories and which reinforces the distinctiveness of sounds that cross categories. The experimental data imply that the representation is suprasegmental, is organized syllabically, and classifies the contextual and dynamical acoustic features of speech in a way which captures the phonetic similarities among sounds.

The generalization of newly mastered sounds to new lexical contexts and the close relationship between children's speech and parental target words suggest that children's utterances are composed from elemental units, even if the resulting compositional structure does not match the phonotactic constraints of adult phonology. Pronunciation errors suggest a contrasting view of children's speech in which anatomical constraints, articulatory complexity, and motor skill play a central role in explaining their evolving phonological competence. Patterns of variation, however, imply that articulatory and compositional components each have an equal role in explaining children's speech production. The dual character of speech suggests a motor control structure which is hierarchical. Several other properties of children's speech production point to associative, stochastic mechanisms that link children's internal representation of speech and their utterances.

4.2 Resolving issues of phonological development

Our resolution of these issues is in two parts — perceptual and motor. We propose a perceptual system which detects and categorizes local and contextual phonetic features of speech, generating a faithful syllabic representation which governs speech recognition and production. We also propose a hierarchical motor control system composed of a phonological control level and an articulatory control level. The

former composes an utterance out of elemental sounds by choosing a sequence of articulatory events most likely to generate each sound, and the latter generates elemental sounds by choosing a sequence of articulatory gestures. The model's auditory perception is specialized to segment and categorize acoustic feedback into discrete phonetic events which closely correspond to discrete sets of gestures learned by the model's articulatory controllers. To imitate sounds, the model need not solve the hard problem of relating continuous speech and continuous vocal tract motion. The phonological control level learns the correspondence between one discrete sequence of phonetic events and a discrete sequence of articulatory events.

A categorical phonetic representation of the sort we propose has additional advantages of communicational robustness and computational compactness. Its compactness and faithfulness enable a more economical memory of utterances and utterance types and the possibility of off-line learning of sounds without parental supervision. Its encoding of contextual features in a whole-syllable representation forms the perceptual foundations necessary to support compositional structure. Yet, the auditory system we propose relies on exclusively acoustic cues and their statistical distribution in the environment.

The two levels of hierarchical motor control operate at very different temporal resolutions. This might be a problem except that the discrete phonetic events detected by auditory perception regulate the timing and decisions of the phonological control level, while continuous proprioceptive feedback and the intrinsic dynamical properties of the vocal tract and respiratory system regulate the timing and decisions of the articulatory control level.

The model uses reinforcement learning to simultaneously learn the kinematics and dynamics of the articulatory system and solve other problems of temporal credit assignment in a relatively economical way. Reinforcement is determined by comparing the phonetic representation of a target utterance with the phonetic representation of the actual utterance, a faithful representation of the relative order of sounds in each syllable, including discontinuities. The model also addresses how to control the vocal tract's many degrees-of-freedom, problems of hierarchical reinforcement learning, and other constraints to reduce the computational complexity of the model's task.

4.3 Model components

Figure 4.1 portrays HABLAR's architecture. Each box represents an implemented computational component of the model. Internal representations and flow of information are indicated by solid lines in the diagram. Acoustic signals are indicated by dashed lines.

HABLAR's speech is a synthetic acoustic signal computed on the basis of lung and vocal tract activity. It is generated by a version of Haskins Laboratories' ASY *articulatory synthesizer* (Rubin et al. 1981), which computes the acoustic properties of the vocal tract, and a *source model* of our own design, which computes intensities of voicing, frication, and aspiration based on respiratory mechanics, vocal tract configuration and aerodynamics. Speech is generated by the model or by a synthetic "parent". The "parent" is merely a set of synthetic utterances presented to the model in ways intended to simulate a human child's linguistic environment.

The model's speech generation requires control of twelve articulatory dimensions involving the position or state of the glottis (the vocal cords), lungs, jaw, tongue body, tongue tip, lips, velum, and hyoid.[1] For each articulatory dimension, there is one *gesture controller* whose activity emulates a smooth skeletal-muscular trajectory. There are many *articulatory controllers*; only one is active at any one time. To generate a component sound, it coordinates multiple articulatory gestures in parallel in all articulatory

[1] The velum is a flap which controls the opening between oral and nasal cavities. The hyoid is a bony structure joining epiglottis, tongue root, and muscles attached to several other structures.

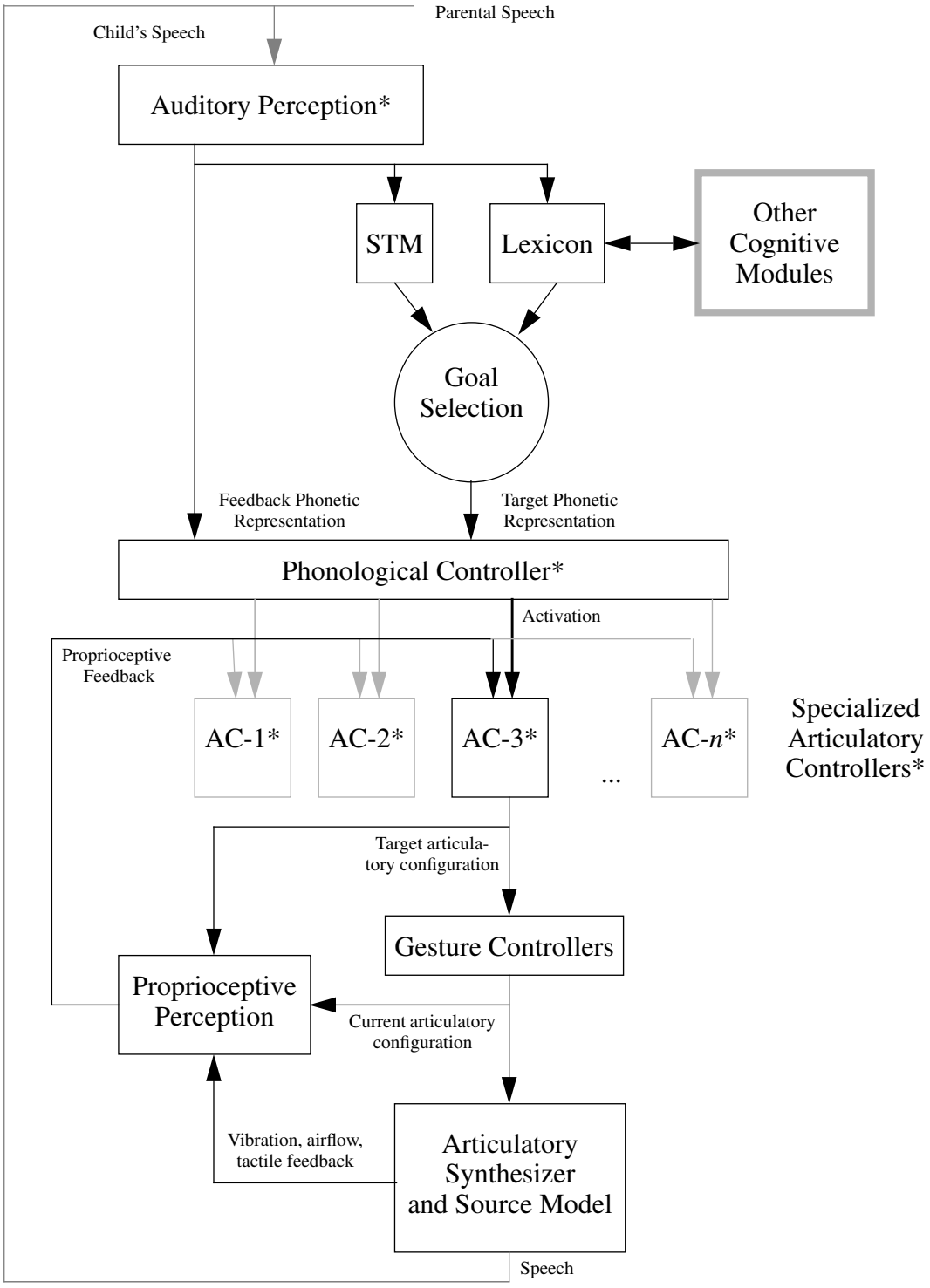


Figure 4.1 HABLAR's cognitive architecture

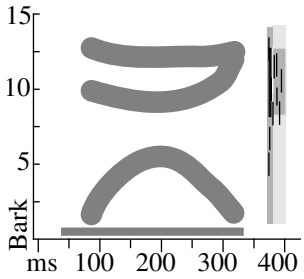
Each box represents an implemented component of the computational model. Modules marked with an asterisk (*) are adaptive, including at least one significant learning component. In *auditory perception,* only phonetic categorization is learned. Flow of information is indicated by solid lines and arrows. Acoustic signals are indicated by dashed lines.

dimensions, observing tactile, kinesthetic, and proprioceptive feedback compiled by *proprioceptive perception* from vocal tract, respiratory system, and gesture controllers. The *phonological controller* coordinates a sequence of articulatory events, activating one articulatory controller after another to compose an entire utterance.

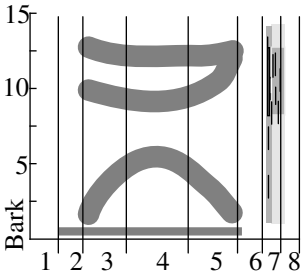
Auditory perception segments speech into a set of discrete phonetic events by dividing it into periods of silence, friction, and voicing and by identifying periods of maximal and minimal spectral change. It sorts static and dynamic spectral patterns during each of these periods into categories determined according to the statistical distribution of sounds observed in parental speech. Phonetic representations of past utterances are stored either in *short-term memory* or in (long term) *lexical memory.*

4.4 Imitating *duck*: A sample scenario

We illustrate how the model might ideally imitate the word *duck* [dAk^{h}]. The episode begins when the model's "parent" speaks, generating a synthetic acoustic signal which is processed by auditory perception to form a phonetic representation. This superimposed activation of all phonetic events detected over the course of the utterance is stored in short-term memory. Omitting some details for greater clarity, phonetic events which are detected and stored correspond to demisyllables [dA] and [Ak], release [k^{h}], and vowel [A]. These are illustrated in Figure 4.2. The goal selector copies the representation from short-term memory to the phonological controller to start imitation.



a. Continuous



b. Segmented

Index	Segment Type	Spectral Category
1	Silence	
2	Prevoicing	
3	Transition	[dA]
4	Static	[A]
5	Transition	[Ak]
6	Silence	
7	Friction	[k6]

c. Phonetic events detected

Figure 4.2 Acoustic signal and phonetic events

Idealized spectrogram of *duck* before (a) and after (b) segmentation. The list of phonetic events (c) tabulates the type and spectral category (when applicable) of each detected segment. The word's phonetic representation is a superpositional vector whose units are incremented each time a feature is detected — each segment type, each triad type (not shown, which captures the relative order of segment types), and each spectral category. Further details may be found in Chapter 6.

Given the target sound (e.g, [dAk^{h}]), the phonological controller's overall goal is to activate the right set of articulatory controllers in the right sequence, thus reproducing the target's phonetic representation. Its first step is to determine which articulatory controller is most likely to accurately reproduce the onset demisyllable [dA]). The chosen articulatory controller is activated. It in turn chooses a sequence of articulatory gestures which generates the sound. Once the lungs are filled, the first set of gestures is chosen. In one gesture, the jaw partially closes; in another, the tongue tip rises to touch to alveolar ridge (the roof of the mouth behind the upper incisors). The active articulatory controller does not choose each point through which the jaw and tongue will move. Rather, for each articulatory dimension, it chooses a target equilibrium point (e.g., jaw angle, tongue tip length and angle at closure). The gesture controller (one for each

articulatory dimension) realizes the motion as a smooth trajectory between the articulator's current position and its new target equilibrium point.

When tongue-alveolar contact is detected, the active articulatory controller chooses one gesture which relaxes inspiratory muscles (to start exhalation) and another which closes the glottis such that vocal cords vibrate as soon as air passes over them during exhalation. The tongue still constricts free air flow. When oral air pressure rises above a threshold, another set of gestures is chosen which opens the jaw, relaxes the tongue tip, and moves the tongue body to the position necessary to pronounce [A].

At each discrete time step, airflow, lung volume, vocal tract and pulmonary configurations, tactile and proprioceptive signals, turbulent and periodic sound amplitude, and vocal spectrum are calculated. Proprioceptive feedback is transmitted to the active articulatory controller. Acoustic feedback is transmitted to the model's auditory perception. The spectral pattern during the onset demisyllable reveals a rising first formant and slightly falling second and third formants. Auditory perception recognizes this rapidly changing spectrum as the onset demisyllable [dA]. However, as the tongue approaches its equilibrium position, its motion slows, and the rate of change in the voice's spectrum drops. Now auditory perception detects and recognizes the spectral pattern corresponding to the vowel [A].

Once [A] is detected, the phonological controller is free to choose a different articulatory specialist to generate [Ak] and yet another for the final [k^h] release until the phonetic representation of the acoustic feedback matches the phonetic representation of the target utterance or until the imitation attempt ends for other reasons (e.g., if the model's lung capacity is exhausted).

This scenario is revisited when motor control is presented in greater detail in Chapter 7 and is illustrated in Figure 7.3. Chapter 5 details speech generation, proprioception, and gestural dynamics. Chapter 6 details auditory perception and phonetic representations.

4.5 The model's operation

The model has two fundamental behaviors: imitation of sounds in the environment or intentional speech. To initiate a behavior, the *goal selector* copies the target sound's phonetic representation from short-term memory (for imitation) or from lexical memory (for intentional speech) to the phonological controller. The model does not "intentionally" babble. Babble is just the result of imitation or intentional speech whose referent is unclear to the listener because of ill-formed sounds (which may occur because of the model's exploratory behavior).

Each articulatory controller is a specialist capable of generating all or part of a recognizable sound. Given a target sound, the phonological controller determines which articulatory controller is most likely to accurately reproduce it. The chosen articulatory controller is activated, and it in turn chooses a sequence of articulatory gestures which generates the sound. Each gesture is a smooth trajectory (resembling the motion of a damped spring) between the articulator's current position and its new target equilibrium point (Browman & Goldstein 1989) as realized by a gesture controller.

At each discrete time step, airflow and lung volume are calculated and vocal tract and pulmonary configurations are updated. The articulatory synthesizer and *source model* convert information on airflow and vocal tract configuration into a synthetic acoustic signal whose feedback is analyzed by auditory perception. The acoustic signal may also be physically realized as synthetic speech for experimental observation of the model's behavior.

The active articulatory controller's decisions are not governed by auditory feedback. Rather, continuous tactile and proprioceptive feedback governs its choice of articulatory gesture in each articulatory dimension. The articulatory controller does not depend on some sort of internal clock for its timing. Rather,

proprioceptive feedback allows the articulatory controller to take advantage of the intrinsic dynamical properties of articulatory gestures and respiration for gesture timing.

Simultaneously, the auditory perception module monitors the acoustic consequences of articulatory controller and gesture controller activity, interpreting it phonetically, and building a cumulative representation of the phonetic events detected over the course of the model's utterance. The phonological controller monitors new phonetic events, choosing when to deactivate the previously chosen articulatory controller and when to activate a different articulatory controller better suited to generate the target utterance's next phonetic event. This continues until the phonetic representation of the acoustic feedback matches the phonetic representation of the target utterance or until the system gives up.

The phonological controller attempts to follow a trajectory of phonetic events which achieves the goal. It does not attempt to control the articulatory configuration directly and under normal circumstances has no knowledge of the articulatory or proprioceptive state; it selects articulatory controllers which can be expected (in a statistical sense) to achieve the phonetic state it desires. The active articulatory controller attempts to traverse a proprioceptive trajectory which optimally generates the sound desired by its phonological client.

4.6 Components that learn

HABLAR's adaptive modules are the auditory system's phonetic categorization processes, the phonological controller, and articulatory controllers.

Auditory perception employs soft competitive learning (Nowlan 1991a,b) to learn categories of static spectra and dynamic spectra by building a statistical model of the static and contextual sound patterns in the linguistic environment. Categories of static spectra correspond to vowels, sonorants, and static fricatives. Categories of dynamic spectra correspond to demisyllables and aspirated consonant releases.

The phonological controller employs an implementation of Q-learning (Watkins 1989) adapted from Lin (1992) to learn an optimal policy that maps target and feedback phonetic representations to a choice of articulatory controller. Articulatory controllers use using a parallel version of Q-learning (Markey 1994a, Chapter 7) to learn a policy linking proprioceptive feedback and gestural targets.

Veridical representations of words stored in lexical memory or of previous parental utterances stored in short-term memory guide training. The phonological controller receives a reward proportional to the proximity of target and feedback phonetic representations. The active articulatory controller receives a reward proportional to the proximity of the phonetic representations of target utterance and that portion of the auditory feedback received while it is active.

4.7 Phased learning

HABLAR's perceptual categories are learned first, before any motor learning. This approximates true developmental timelines (see Section 1.1). Anatomical constraints prevent true speech sounds before three months (Kent & Murray 1982), and babbling does not show adult characteristics until at least six months (Oller & Lynch 1992). By this time children already recognize categories of vowels in their native tongue. Only two months later, while they are still exploring consonant-vowel syllables in their babble, children can recognize adult categories of consonant-vowel syllables. It is not until the end of the first year that consonants in their babble mimic the linguistic environment (Vihman et al. 1986).

Strictly speaking, perceptual and articulatory milestones overlap in children. Strict phasing of perceptual and articulatory training probably simplifies the model's developmental task. More natural learning schedules are beyond the present scope of the research.

The computational difficulty of hierarchical reinforcement learning (Singh 1992b,c) suggests the introduction of "bootstrapping" strategies which gradually increment phonetic complexity of target utterances and which gradually phase articulatory and phonological controller training. To implement these strategies, simulations vary the order of stimulus presentation or artificially control the rate and relative timing of learning in articulatory and phonological controllers (see Chapter 7). Some of these heuristics are designed to resemble the variety of idiosyncratic developmental strategies which children seem to employ — avoidance of apparently hard-to-pronounce classes of sounds, preferences for other, less difficult sounds, and analytic strategies (Menn 1983).

4.8 Built-in knowledge and components

Domain-specific knowledge is incorporated into HABLAR in a number of ways. There are non-cognitive components. The articulatory synthesizer specifies the anatomy and acoustic properties of the vocal tract. The source model specifies mechanical, dynamic, and sound-producing aerodynamics (voicing, frication, and aspiration) of the respiratory system and vocal tract. Dynamical properties of the muscular-skeletal system and gestural organization of articulatory motion are specified by the gesture controllers.

Except for categorization of static and dynamic spectral features, most perceptual facilities are assumed to be either innate or learned prior to the start of speech-like babbling. This includes tactile, kinesthetic, and proprioceptive perception, plus categories of discrete tactile events. It also includes auditory mechanisms of basic signal processing, spectral analysis, cues used for acoustic segmentation, segmental features, and syllable structure features. Certainly, categories of tactile events, segmental features, and syllable structure features might be learned, but how they are learned is beyond the scope of this research. We assume them as prior knowledge, remaining agnostic about their origin.

Specific details of several high-level cognitive faculties are beyond the scope of this research. Short-term memory, lexical memory, and goal selection are specified procedurally. The contents of lexical memory are a function of experimental manipulations used to test the model's behavior under different conditions.

Chapter 5 Articulatory Anchors and Speech Generation

We begin a detailed description of *HABLAR* with its peripherals. They constrain and anchor the model. Laying the groundwork now will make it easier to explain other parts of the model later.

5.1 Speech generation by the articulatory synthesizer and source model

The synthetic acoustic output of child and parent is generated by an adaptation of Haskins Laboratories' ASY articulatory synthesizer (Rubin et al. 1981, see also Mermelstein 1973). At each discrete articulatory controller time step (corresponding to 8 msec of real-world time), it converts a static vocal tract configuration into a synthetic acoustic signal. By animating a trajectory of articulatory configurations, the model simulates the dynamic acoustics of speech.

Each vocal tract configuration is defined by control parameters for six articulators which vary over 10 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). Figure 5.1 displays a midsagittal profile of ASY's vocal tract and the location of various articulators and other anatomical features. To ASY we have added a *source model*, a module which computes voicing and friction source intensities and their location in the vocal tract based on pulmonary mechanics (Levitzky 1991, Mines 1993) and a model of vocal tract aerodynamics (see Stevens 1971). These are a function of vocal tract configuration and two additional degrees of freedom: *glottal opening* (distance) and *extra-pulmonary pressure* (cm H_20) generated by inspiratory muscles. Together, vocal tract and source models have 12 degrees of freedom.

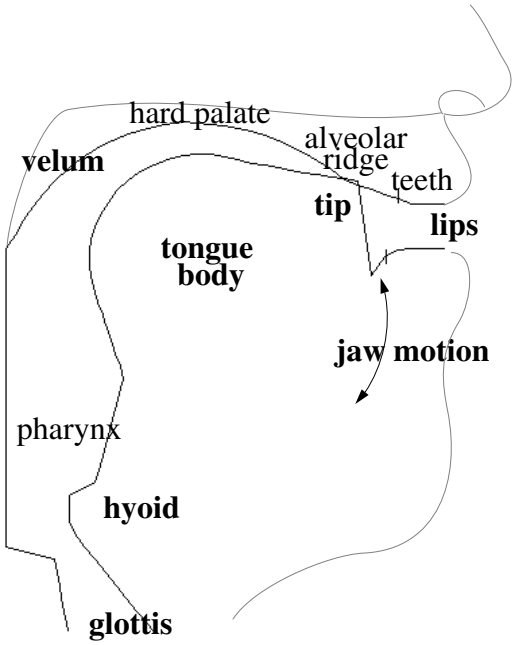


Figure 5.1 Vocal tract articulators

Solid outline is drawn by ASY. Control dimension labels are bold. Other outlines and labels are added for context.

From these input parameters, ASY computes the coordinates of each articulator, a midsagittal 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 HABLAR's synthetic acoustic output: a power spectrum, voice and friction amplitudes. Synthesized sound and graphics of the vocal tract are available for experimental observation of HABLAR's behavior. Based on airflow, glottal vibration, articulator coordinates, midsagittal outline, and cross-sectional areas, the model also simulates tactile sensation.

5.2 Gestural organization of articulatory motions

Several investigators have proposed that humans produce speech as a sequence of articulatory gestures as if they were notes in a musical score (Browman & Goldstein 1989; see also Kelso et al. 1986, Saltzman & Munhall 1989). Some evidence also suggests that gestural control of articulatory motion is well formed, even among infants as young as 5 months (Piroli 1991). Thus, changes in vocal tract and pulmonary configurations are specified 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 smooth trajectory conforming to the motion of a critically damped spring (Browman & Goldstein 1989, see also Polit & Bizzi 1978). 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. A *gesture controller* realizes the actual trajectory for each vocal tract and pulmonary articulator.

Gesture trajectories are updated at each discrete time step (8 msec of real world time) in accord with the equation of motion for a critically damped system:

$$\ddot{x} + 2\beta\dot{x} + \beta^2 (x - x_0) \;=\; 0 \tag{5.1}$$

where β is a damping coefficient determined by articulator mass m and spring constant k according to $\beta^2 = k/m$, x is displacement, and x_0 is displacement from equilibrium at time 0. The closed form solution is

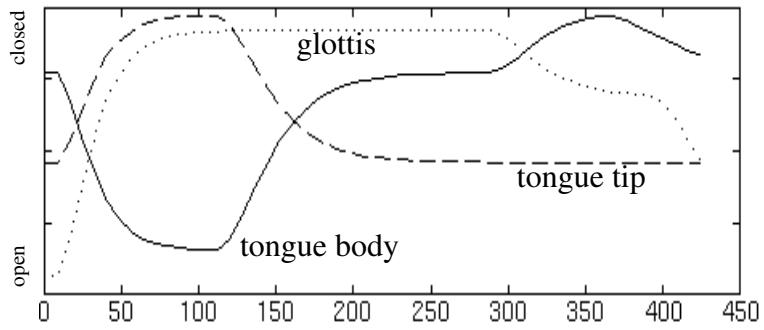
$$x(t) \;=\; x_0 + (\beta x_0 + v_0)\, te^{-\beta t} \tag{5.2}$$

where v_0 is velocity at time 0, and t is 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.

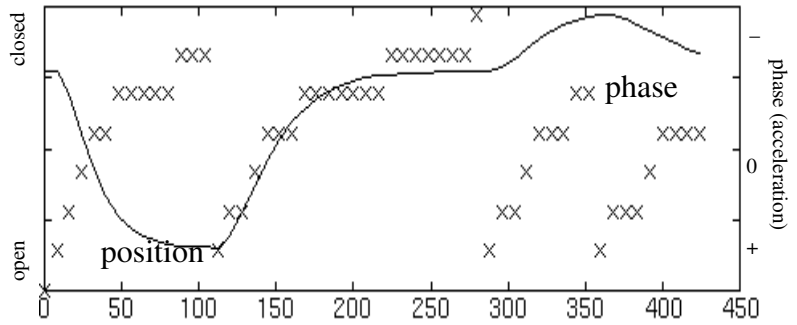
HABLAR has no clock with which to time motor control events. Instead, it takes advantage of the intrinsic dynamics of vocal tract articulators as realized by gesture controllers and as revealed by the timing of proprioceptive feedback (see Section 5.3).

To augment proprioceptive signals, the gesture controller also tracks each gesture's progress with a discrete measure which divides its ballistic trajectory into several parts. Although the trajectory is not periodic, we call this measure "phase", because it roughly corresponds to the "phase" of a periodic or semi-periodic dynamical system in state space. When an articulator is at rest, it has a *phase* of zero; during its initial acceleration, it has a phase of 1; after initial acceleration but before the articulator starts to decelerate, the articulator has a phase of 2. A phase of 3 signals that acceleration has changed sign and absolute velocity has reached its maximum. Then, as the articulatory decelerates, its phase increases from 4 to 7 until it comes to rest, and phase is reset to zero. Phase is computed as a simple consequence of gesture execution.

Tongue body, tip, and glottis trajectories generated by a parental gesture script for *duck* are illustrated in Figure 5.2a. The tongue tip closes at about 100 msec for the /d/. The tongue body closes at about 350 msec for the /k/. The glottis, initially wide open, is closed only enough for voicing to occur, but is opened at about 300 msec for the voiceless /k/. The graph in Figure 5.2b indicates position and gestural phase of the tongue body. Position is identical to tongue body position shown in Figure 5.2a. Phase is shown as cross-hatches, indicating four separate tongue body gestures. Starting at rest (0 msec), the tongue body at first moves forward to assist the tongue tip's contact during the /d/ with the alveolar ridge (shown here as a relatively "open" position and also shown in the vocal tract profile). After alveolar contact is established and once oral pressure reaches a threshold (112 msec), the stop is released into the vowel for the second gesture. The third gesture begins after the tongue body has sufficiently decelerated and phase exceeds 6 (at 288 msec). The vowel is then complete. Now the tongue body moves to form a velar closure corresponding to the /k/. Once closure is detected (360 msec), the fourth and final gesture releases the /k/.



a. Several articulatory gesture trajectories during *duck*



b. Tongue body position (solid curve) and phase (cross-hatches) during *duck*

Figure 5.2 Articulatory gestures

5.3 Proprioceptive perception

Proprioceptive perception recognizes tactile stimuli and proprioceptive events, compiling the state and history of vocal tract and pulmonary activity from a number of sources. The resulting analysis becomes the proprioceptive state, common input for each articulatory controller. A total of 118 units encode the proprioceptive state in the model's current implementation. A summary of the representation follows.

- **Current vocal tract configuration** is a real vector of articulator positions, the outputs of gestural controllers, each scaled between 0 and 1.

- **Target vocal tract configuration** is a real vector of target articulator equilibrium positions, the outputs of articulatory controllers, each scaled between 0 and 1. This is not strictly proprioceptive but together with current vocal tract configuration defines an approximate trajectory.

- **Gestural phase.** Phase is 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 distributed cyclic 6-unit representation starting with zero for "at rest" and proceeding from phase 1 through 7 thence again to zero.

- **Touch.** The model detects touch and the recent history of articulatory contact at regions in the vocal tract roughly corresponding to places of articulation (glottal, pharyngeal, laryngeal, uvular, velar, palatal, postalveolar, alveolar, dental, labial). For each region, 1 represents current contact between opposing articulators; 0 represents no recent contact. After contact is released, the value decays exponentially with a half-life of 52 msec. In addition, a single binary unit represents whether closure is detected anywhere in the oral tract.

- **Vibration and Turbulent Flow.** Global sensors for current and recent glottal vibration and turbulent airflow are also included. A single binary unit detects current glottal vibration. Values of two units which represent current vibration and turbulence are scaled with an inverse exponential function which ranges between 0 (for no present vibration or turbulence) and 1. Present and recent vibration and turbulence is represented by five units which encode five different conditions: vibration, oral turbulence, glottal turbulence, no vibration, and no turbulence. Each is set to 1 when its condition is true and decays exponentially after the condition becomes false with a half-life of 52 msec. There is little evidence of tactile feedback for specific locations of turbulent flow (except labial frication). However, a global feedback signal of the sort proposed here is justified if based on non-glottal airflow resistance.

- **Respiration.** Various pulmonary state measures are represented, including lung volume, airflow (glottal, oral, and nasal), and air pressure (subglottal, oral). Airflows are scaled between -1 and 1 using the hyperbolic tangent. Zero represents no airflow. Lung volume is scaled likewise, measured relative to the respiratory system's equilibrium (functional residual) volume.

5.4 Source model implementation details

Haskins Laboratories' ASY articulatory synthesizer does not compute vocalic or turbulent sound amplitudes from a respiratory or glottal model. Designed as an interactive tool, it depends on the experimentalist to provide sound amplitudes as inputs. Thus, we have implemented a simple model of vocalic and turbulent sound generation. A complete model is too computationally expensive for the purposes of the present research (e.g., Flanagan et al. 1975). The model we implemented is grounded in pulmonary mechanics and simple aerodynamic models of voicing, frication, and aspiration. It calculates voicing and friction source intensities and their location in the vocal tract as a function of vocal tract configuration (as computed by the articulatory synthesizer), glottal opening, and extra-pulmonary pressure generated by inspiratory muscles.

Respiration proceeds as a cycle of inspiration and expiration. During inspiration, the diaphragm and chest wall muscles expand the chest cavity and air flows into the lungs. During expiration, these muscles relax, which induces airflow to restore the equilibrium between the pressure of collapsing lungs P_l and the chest wall P_w. At equilibrium, there is no airflow U and subglottal pressure P_s is zero. During inspiration, subglottal pressure is negative; during expiration it is positive. The relationship among subglottal pressure P_s, the effective pressure of the collapsing lung P_l, the opposing expansion of the chest wall P_w, and the effective pressure of inspiratory muscles P_m is expressed as

$$P_s = P_m + P_l + P_w \tag{5.3}$$

where P_l and P_w are functions of lung volume V and the respective compliances of lung and chest wall (Levitzky 1991, Mines 1993).

Opposing the subglottal pressure is friction due to airflow through constrictions of the glottis and vocal tract. The resistive pressure drop P_r due to turbulent flow is:

$$P_r = \frac{k\rho U^2}{2A^2}, \tag{5.4}$$

where k is a constant, ρ is the air density, U is the volume velocity (airflow), and A is the cross-sectional area of the constriction (Stevens, 1971). Resistive pressure drop is distributed among glottal and vocal tract constrictions, added directly if constrictions are in series and reciprocally if they are in parallel (Levitzky 1991).

$$\begin{aligned} P_{\text{oral}} &= \sum_{i \in \text{ constrictions}} P_i \\ \frac{1}{P_{\text{supra}}} &= \frac{1}{P_{\text{nasal}}} + \frac{1}{P_{\text{oral}}} \\ P_s &= P_{\text{glottal}} + P_{\text{supra}} \end{aligned} \tag{5.5}$$

Together, these relationships define a dynamical system for which there is no closed form expression. Therefore, its behavior is estimated by a simple interpolatory numeric integration method. At each discrete time step, the configuration of the vocal tract, glottis, and the current extra-pulmonary pressure due to inspiratory muscles P_m are computed.

Then P_l and P_w are determined based on the current lung volume V, and P_s is computed per (5.3). The next step is to determine volume velocity at the glottis and to update the lung's volume. But glottal airflow is a function of P_s and the effective cross-sectional area of the entire vocal tract, which can be determined from (5.4) and (5.5). Thus, we first determine effective cross-sectional area A_s from glottal, nasal, and oral cross-sectional areas as follows:

$$\frac{1}{A^2_{\text{oral}}} = \sum_{i \in \text{ constrictions}} \frac{1}{A^2_i}$$

$$A^2_{\text{supra}} = A^2_{\text{nasal}} + A^2_{\text{oral}} \tag{5.6}$$

$$\frac{1}{A^2_s} = \frac{1}{A^2_{\text{glottal}}} + \frac{1}{A^2_{\text{supra}}}$$

and then solve (5.4) for U and integrate between the previous and current time step to update V. A small correction is made which accounts for the small elasticity of the oral tract during closure. Once the system's volume velocity is determined, pressure drop at each constriction is backed out using (5.4) and (5.5). Turbulent sound pressures, *Fric*, are roughly a function of cross-sectional area and pressure drop at each constriction (Stevens 1971), and vocalic sound pressure, *Voice,* due to glottal vibration is roughly a function of glottal pressure drop, as follows:

$$\begin{aligned} Fric_i &= K_F \sqrt{P_i^3 / A_i} \\ Voice &= K_V P_{\text{glottal}} \end{aligned} \tag{5.7}$$

These relations are approximate. Glottal vibration (voicing), for example, does not occur below lower and upper limiting cross-sectional areas, each of which depends on glottal pressure drop (Flanagan et al. 1975). Voice and frictional thresholds (Stevens 1971) are applied, attenuating *Fric* and *Voice* at their respective boundaries.

Chapter 6 Auditory Perception

This chapter describes HABLAR's perceptual foundations. The model's auditory perception module is designed to satisfy empirical and modeling constraints identified in previous chapters. It generates a veridical representation of adult speech or the acoustic feedback of HABLAR's own speech which governs the pronunciation of target utterances. The representation simulates the learned categorical phonetic perception of late infancy in which acoustically similar sounds are assimilated into existing categories (e.g.Kuhl et al. 1992, Werker et al. 1981). The representation is suprasegmental. HABLAR's auditory perception is sensitive to contextual and dynamical acoustic properties, syllable structure, and organizes speech sounds syllabically (e.g., Jusczyk et al. 1995). Though categorical perception is learned, HABLAR does not attempt to simulate the sequence of events which leads from acoustic to categorical phonetic discrimination, nor does it address the larger issues of word recognition and lexical access (but see Jusczyk 1993).

6.1 Motivation and Background

Human speech and human listening presumably evolved together. It is therefore plausible that speech perception is specialized not only for word and sentence recognition but also for the unique demands of articulatory motor control and that it plays an important role in articulatory and phonological development. Perception's role in motor control is seldom acknowledged except as the source of target sounds to be imitated or learned. However, milestones of perceptual development always precede corresponding milestones of motor development. We hypothesize that the categorical character of auditory perception which underlies robust word recognition also acts as a grammar which defines the well-formedness of children's speech, shaping the distribution of sounds in their productive repertories. The connection is deeper: we conjecture that without the ability to parse speech into discrete perceptual events, learning to speak would be difficult, and it would not be possible to account for the compositional structure of speech which emerges in childhood.

Importantly, the perceptual model relies on exclusively acoustic cues and their statistical distribution in the child's linguistic environment; it avoids prior assumptions of articulatory-acoustic correlations or linguistic contrasts. It is inappropriate to model perceptual development with features which assume prior knowledge of articulatory-acoustic correlations or semantic contrasts, knowledge the prelinguistic infant does not possess (see Section 3.3). Segments and categories the model detects are not the mature phonemes of adulthood; nor do they correspond to distinctive features (Chomsky & Halle 1968), traditional articulatory features (Ladefoged, 1972), or any of myriad alternatives (e.g., Shillcock et al., 1992). The model's acoustic segments are longer in duration than phonemes or phonological features, at least long enough to capture coarticulation between contiguous consonants and vowels and to detect features of syllable structure. They correspond to coarse-grained changes in voicing, friction, and spectra. They identify a syllable's most salient spectral features.

Although such segments and categories are based on exclusively acoustic measures, they correlate with linguistic events and delineate articulatory gestures. The model's auditory perception is specialized to segment and categorize acoustic feedback into discrete phonetic events which closely correspond to discrete gestures learned by the articulatory apparatus, but to do so based on minimal assumptions.

6.2 An acoustic-based phonetic representation

Auditory perception's *input* is an unsegmented acoustic representation of parental speech or feedback from the model's own speech. Its *output* is a phonetic representation of the sequence of acoustic segments and spectral categories detected in the utterance. We use *duck* as an example to introduce how the model segments and categorizes speech parcels. Figure 6.1a illustrates the model's continuous acoustic input as a schematic spectrogram of *duck*. Its vertical axis is frequency (using the "Bark" scale), the horizontal axis is time, and dark patches represent high acoustic energy. The thick dark bands represent formants, the changing resonant frequencies of the vocal tract; the variously-shaded vertical area near 450 msec is a burst of noise. The horizontal bar near zero frequency represents the underlying vibration of the vocal cords.

A reasonable first step in parsing the unsegmented acoustic signal is to divide it into broadly classified periods of sound — continuous periods of silence, voicing, and friction (aspiration or frication). This approach yields a period of voicing between 40 and 330 msec, friction between 370 and 410 msec, and three periods of silence. This is plausible, given the broad segmentation of speech patterns by the auditory nerve's adaptation properties (Seneff, 1988), but it is clearly not sufficient to identify linguistically relevant spectral features.

To further divide the utterance, we consider locating temporally stable spectral patterns associated with the steady-state portion of vowels and other sonorants, fricatives, and aspiration. Consonants are the result of the vocal tract in motion to or from a vocal tract configuration (e.g., closure) whose acoustic properties are not entirely discernible or discriminable. Even the spectral pattern which accompanies some vocal tract closures, such as the nasal undertone which accompanies /m/, /n/, and /N/, does not reveal the stop consonant's place-of-articulation. Consequently, stop consonant identity can be determined only indirectly from the consonant's acoustic trajectory, i.e., from *dynamic* spectral patterns. Points of maximum spectral change are the most salient portion of an utterance for consonant and syllable perception in adults (Furui, 1986; see also Lindblom & Studdert-Kennedy 1967) and are especially salient for young children (Nittrouer & Studdert-Kennedy, 1987; Nittrouer, 1992). Likewise, the salient characteristic of a diphthong is the transition from one vowel to the next. Thus, a reasonable second step in parsing this signal is to identify periods of maximal and minimal spectral change.

By this method, there are eight acoustic segments in *duck*. Their location in the utterance is portrayed in Figure 6.1b, which divides the time axis by segment rather than equal units of time. After an initial *silence* (segment 1), a period of *prevoicing* (segment 2: 50 msec) is detected during which vocal-cord vibration is audible, but the spectrum of the vocal tract's resonant sound structure is obscured by the /d/'s closure.[1] Once a spectrum becomes apparent, it is scanned for the relative degree of spectral change. *Transitions* (3, 5 at 100 and 300 msec) are segments corresponding to maximal spectral change; formant slopes are greatest. The *static segment* (4: 200 msec) corresponds to a period of minimal spectral change; formants are relatively flat. After a period of silence (6: 350 msec) during the unvoiced /k/'s closure, a *burst* (7: 370 msec) of intense but rapidly decaying friction is detected when the contact of tongue body and velum is released. This is followed by a final period of silence (8).

Once segmented, the next step is to draw a sample static spectrum from each static segment and match it with prototype categories of steady-state sounds. Likewise, the model samples dynamic spectral properties during transition or burst segments and matches the sample against prototype categories of

---

[1] Prevoicing is present in other languages, but is not regularly observed in English, except possibly in deliberate speech, in words emphasized in child-directed speech, or in children's speech.

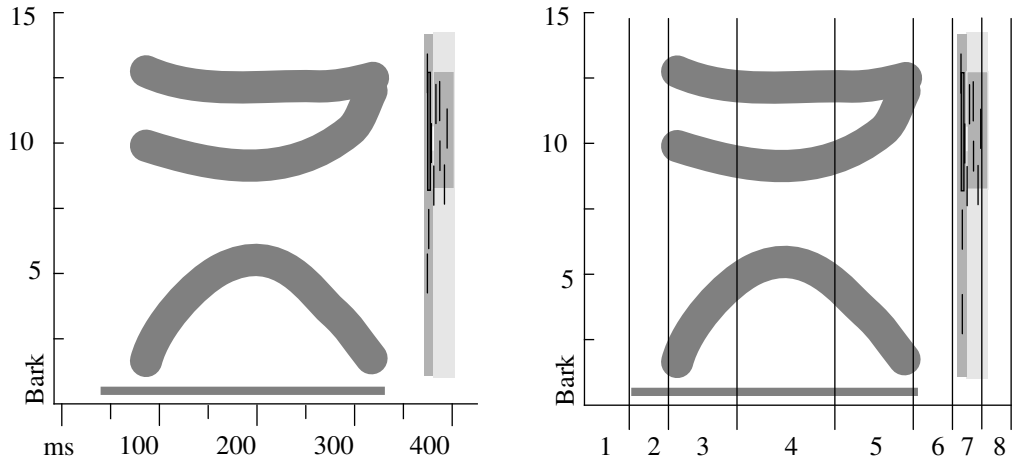


Figure 6.1 Idealized spectrogram of *duck* before and after segmentation

Segments are numbered as follows: (1, 6, 8) Silence, (2) Prevoicing, (3, 5) Transitions [dA] and [Ak], (4) Static segment [A], and (7) Burst [k^h].

dynamic sounds. The model learns an inventory of prototype categories for static and transition spectra from its linguistic environment. In the example above, let us assume that the model has already learned an inventory of prototype spectral categories. Then static spectral properties of segment 4 match the prototype static spectrum corresponding to the vowel [A]. The dynamic spectral properties of segments 3, 5, and 7 match transition spectrum prototypes corresponding to demisyllables [dA] and [Ak], and aspirated stop consonant release [k^h] respectively.

6.2.1 Form of the phonetic representation

This process generates information about the type, absolute order, and spectral properties of each acoustic segment. Other requirements of the complete sensorimotor model such as motor control, lexical access, and short term memory place additional constraints on the optimal form of the model's phonetic representation. They demand a representation whose size does not vary with the length of the utterance (i.e., a representation with a fixed number of features or units), which encodes relative order of acoustic segments, and which nonetheless faithfully captures important acoustic properties of the entire utterance and admits an accurate metric and efficient algorithm for determining the phonetic distance among utterances.

The model employs a phonetic feature vector with one unit for each feature. A unit's activation is increased each time the property it encodes is detected in an utterance, by a degree inversely related to the acoustic distance between speech token and feature prototype.

For purposes of discussion below, we distinguish between two classes of features. *Spectral features*, which we have already alluded to, refer to those units which encode spectral properties and their relative order in the utterance. *Segmental features* refer to those units which encode a second type of information: acoustic segment properties, and their relative order in the utterance. Thus, each class is further divided among noncontextual and contextual feature types.

6.2.2 Spectral features

Prototype categories of static and transition spectra provide the raw material for spectral features. Activations corresponding to prototypes of [dA], [A], [Ak], and [k^h] are increased as each segment is detected in *duck*.

Transition and static segment categories implicitly encode spectral properties and their relative order. For example, in *duck* the [dA] segment encodes the spectral change which occurs as the tongue moves from an alveolar closure for /d/ or /t/ to the opening for the intended vowel /A/. Static segment [A] must follow transition segment [dA]. Segment [Ak] encodes the spectral change which occurs as the tongue makes contact with the velum for the /k/. It cannot precede [dA] except in a multisyllabic utterance.

6.2.3 Segmental features

Acoustic segment classifications and a contextual encoding of segmental order are the ingredients of segmental features. Segment type is distinguished by five acoustic cues — voicing, friction, spectral change, spectral stability, or moments of relatively intense, quickly changing sounds. Eight segment types observed in English and their defining acoustic cues are described in Table 6.1. As *duck* is perceived, segment type activations are increased as each segment is encountered.

Table 6.1 Segment types and defining acoustic cues

Segment type	Voice	Friction	Transient	Stable	Unstable
Silence (0)	0	0	0	0	0
Prevoicing (P)	1	0	0	0	0
Static (S)	1	0	0	1	0
Transition (T)	1	0	1	0	0
Voiced friction (Z)	1	1	0	1	0
Voiceless friction (F)	0	1	0	1	0
Transition friction (X)	0	1	1	0	0
Burst (B)	0	1	1	0	1

The model encodes the relative, not absolute, order of segment types. It does so by forming a cluster of the three most recently detected acoustic segment types. This coding scheme is similar to Wickelgren's (1969) context-dependent allophone sequence encoding. Many segment type clusters encode important linguistic cues. Contiguous prevoicing and transition segments (PT or TP) indicate a voiced stop consonant in many contexts (e.g., the initial consonant in *duck*); a transition-silence-burst subsequence (T0B) signals a word-final unvoiced stop (e.g., the final consonant in *duck*); a transition-static-transition (TST) cluster usually signals a consonant-vowel-consonant syllable; and a static-transition-static (STS) cluster signals a diphthong.

Because of the role of the three-"segment" clusters in encoding the temporal patterns of voicing, friction, silence, and syllable structure, we call them *prosodic triads*. "Prosodic" is used here in the Firthian sense to refer to suprasegmental structural features which capture the canonical profile of the syllable.

6.2.4 Superimposed activations and phonetic distance

To summarize, there are four types of phonetic features. Segmental features include acoustic segment types and prosodic triad types; spectral features include categories of static spectra and transition spectra. Static and transition categories are learned from the linguistic environment. Segmental features are

prior knowledge, either innate or learned at an age earlier than that which HABLAR models (equivalent to about 5-6 months).

The phonetic representation is an activity pattern over a vector of units, one for each feature. Activations accumulate until the end of an utterance. This superposition of contextual and global phonetic features is a faithful representation (Smolensky 1990) of any one-syllable utterance.

6.2.5 Phonetic categorization and spectral category learning

Spectral features are learned. The model builds an inventory of prototype transition and static spectra which represent those transition and static spectra recognized as relatively distinctive according to their statistical distribution in parental speech. Transition spectra and static spectra are classified separately.

Categories are learned and token spectra are classified using soft competitive learning (Nowlan 1991a,b). Soft competitive learning models a set of observations (whether static or transition spectra) as if it were generated by a set of Gaussian distributions. Figure 6.2 illustrates how clusters of observations (in a 2-dimensional plane) might be modeled as a set of 2-dimensional Gaussians (e.g., vowels in a space defined by first and second formants).

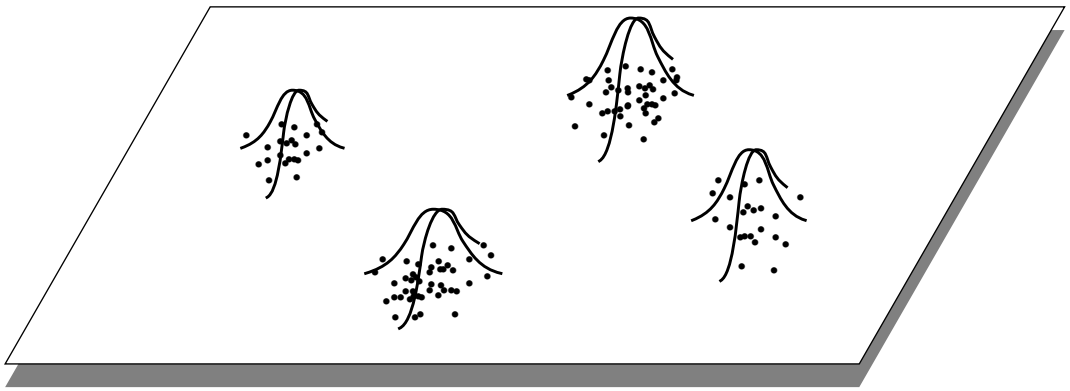


Figure 6.2 Modeling clusters of observations as 2-dimensional Gaussians

Each Gaussian is represented by a mean (a vector of the same dimension as the input patterns) and a covariance matrix. Typically this is simplified by assuming a fixed spherical distribution. Given an observation, the conditional probability that each Gaussian generated the observation is calculated. Each Gaussian is updated by changing its mean and covariance matrix in proportion to the likelihood that it generated the observation.

6.2.6 Redundancy between noncontextual and contextual features

Linguistic theory usually prefers minimally redundant representations. However, there is some redundancy between noncontextual and contextual features, whether spectral or segmental. For example, the dynamic spectral feature corresponding to [dA] and the static spectral feature corresponding to [A] each share information about the vowel. Having chosen to base auditory segmentation and categorization on strictly acoustic data, some redundancy is unavoidable, because there is no possible noncontextual acoustic representation of consonants or diphthongs (see Section 2.8.3), and because static fricatives and sonorants exist in isolation.

In a computational model, information necessary and sufficient to carry out a computation is distributed between algorithm and representation. The less information encoded in the algorithm, the more information must be encoded in data or the data structure. In practice, this is not really different from linguistic theory. In a linguistic theory, redundancy of representations is often minimized by adding rules or introducing default values — the theory's algorithmic components. For example, in underspecification theory (Archangeli 1984, see also Durand 1990), unspecified values of phonetic features (missing data) are filled by default values or according to feature spreading rules.

HABLAR's proposed phonetic representation is minimalist in the sense that it is the most compact form achievable given the model's minimalist algorithmic specification. Neither the model's architecture nor the form of phonetic representations encodes absolute order of phonetic events. The model does not include phonotactic rules. Instead, order is encoded in contextual features.

6.3 Auditory perception implementation details

Auditory perception samples synthetic adult or child sound stimuli once every 8 msec for periodic sound (voicing) amplitude, aperiodic sound (frication or aspiration) amplitude, and 256-frequency power spectrum. Spectral analysis converts the power spectrum from Hertz to Bark, performs Gaussian smoothing over time, and normalizes for total amplitude by computing a difference spectrum. The momentary transition spectrum is the first derivative of the difference spectrum with respect to time. The transition magnitude is the L_1 norm of the transition spectrum minus the L_1 norm of the difference spectrum. Segments are located and segment type is determined based on patterns of periodic amplitude, aperiodic amplitude, and the transition magnitude. Segment types are concatenated to form prosodic triads. Depending on segment type, either a static or a transition spectrum is sampled. A transition spectrum is sampled as an average of the momentary transition spectra for a 50 msec period centered at the point when the transition magnitude is at a local maximum. A static spectrum is sampled as an average of the momentary static spectra for a 50 msec period centered at the point when the transition magnitude is at a local minimum. Once sampled, the static or transition spectrum is categorized, and category units are incremented by the activation of each static or transition category respectively. The information flow of this process is illustrated in Figure 6.3.

6.3.1 Computational resources

The phonetic representation is relatively compact, activating at most a few phonetic features for each identifiable sound segment. Though this substantially reduces the number of parameters which represent the incoming acoustic signal, it is not as compact as a linguistic representational system of phonemes or distinctive features. More than 6,000 real numbers are needed to encode the acoustic input for a 200 msec utterance (about one syllable). Only about a dozen phonetic features are activated by the same utterance. But to encode any adult English syllable, an array of more than 400 phonetic categories are necessary. This takes into account nasals, four fricative places of articulation, four approximants, and twelve vowels (including those which only occur as parts of diphthongs). There are a total of 20 possible static spectral categories in English, plus 330 dynamic spectral categories given the possible consonant-vowel, vowel-consonant, and diphthong combinations. This estimate does not account for combinations which overlap or variations due to coarticulatory effects. In addition, about 50 prosodic triad categories are needed to describe sounds in English. More phonetic categories than traditional phonemes are necessary because our feature system encodes dynamic relationships which extend over several segments. A somewhat smaller representation is used in the model's present implementation, omitting spectral categories corresponding to nasals and fricatives.

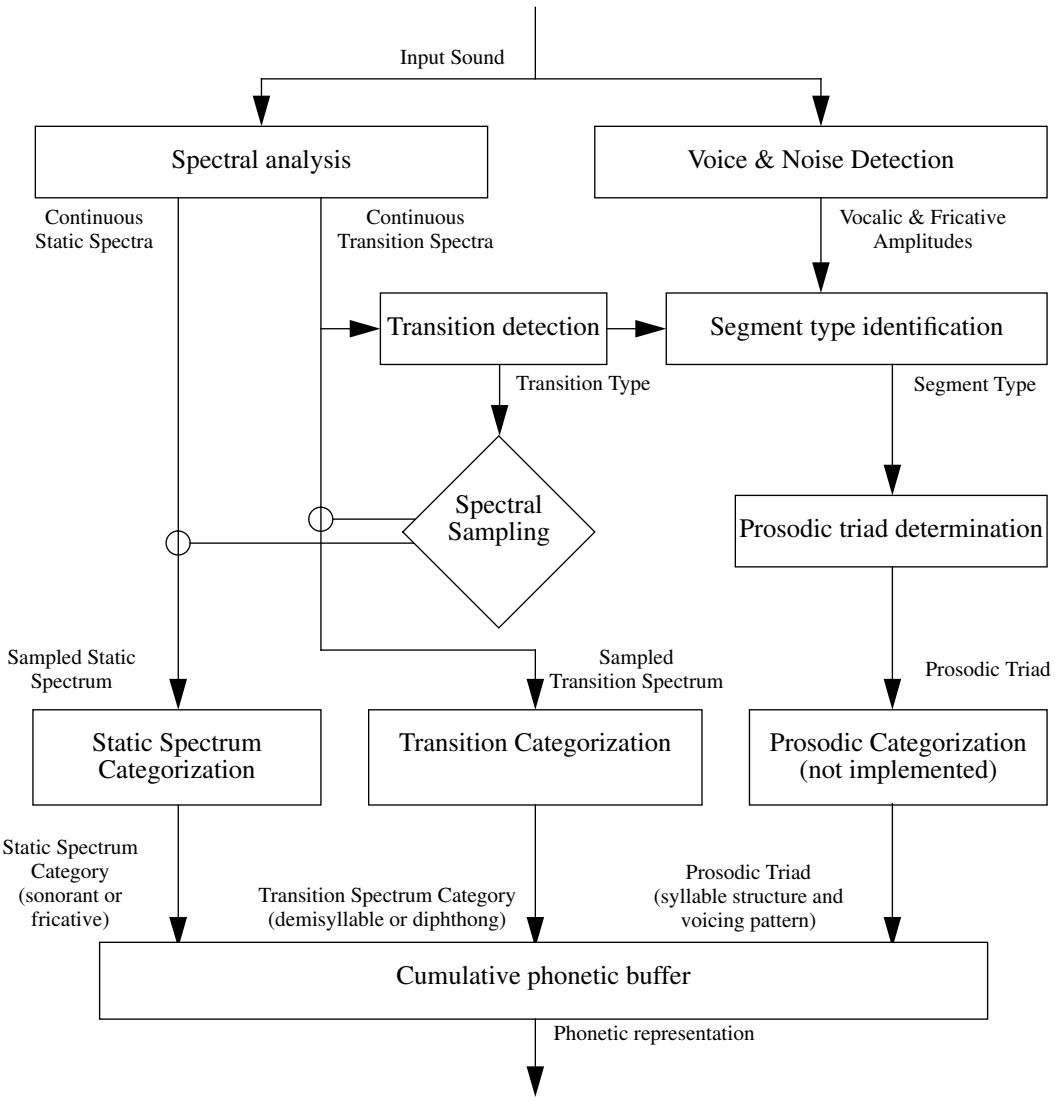


Figure 6.3 Information flow in auditory perception

6.3.2 Spectral analysis

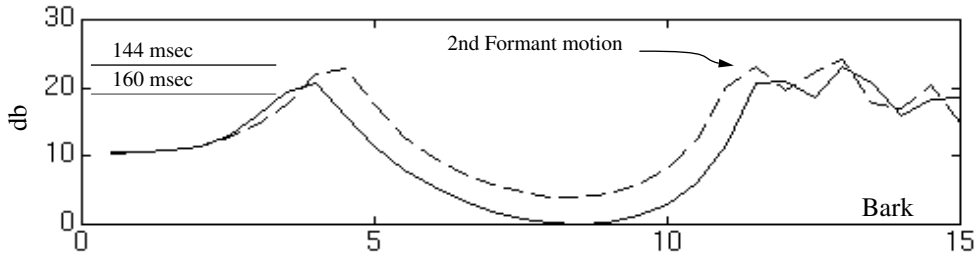
To start spectral analysis, the power spectrum is sampled evenly over the frequency domain. Steps in spectral analysis are shown in Figure 6.5, including representative static, difference, and transition spectra at 144 and 160 msec, peak of the spectral transition corresponding to the onset demisyllable of *duck*.

- **Static Bark spectrum:** To better approximate the frequency response of the human ear and to reduce the size of the parameter space, the model converts frequencies to bark (Zwicker 1961, see also Schroeder et al. 1979) and divides the input spectrum among 44 coarsely coded 1.0-bark wide detectors, b_i for $1 \le i \le 44$, evenly spaced one per 0.5 bark between 0.5 and 22.0 bark.

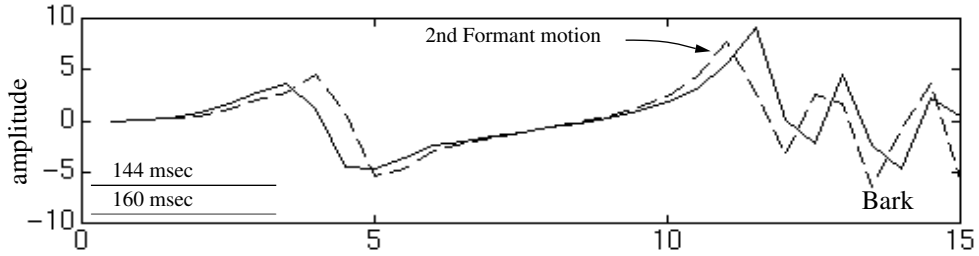
- **Smoothed static spectrum:** To filter short-term noise, the model applies a Gaussian filter (standard deviation of 16 msec) in the temporal dimension (Bradshaw & Bell 1991).

- **Static difference spectrum:** To measure detector amplitudes on a relative, not absolute, scale, the model generates a "difference" spectrum **D**, as the difference in amplitude between neighboring bark detectors, $D_i = b_i - b_{i-1}$ for $i > 1$ (Bradshaw & Bell 1991).

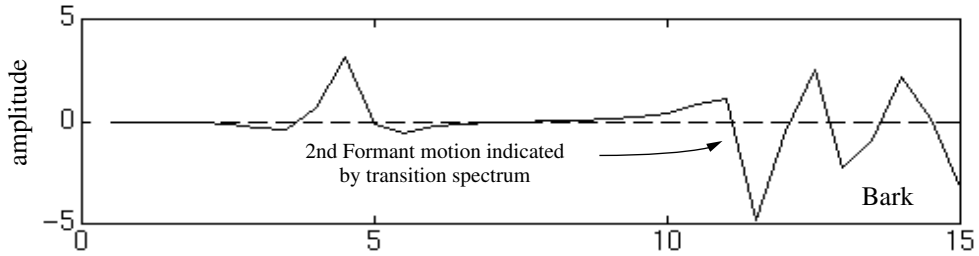
- **Transition spectrum:** To approximate the first time derivative of each difference detector's amplitude, the model computes the difference between detector amplitudes at times t and t–1. The resulting transition spectrum **T** is simply $T_i(t) = D_i(t) - D_i(t-1)$.



a. Smoothed static Bark spectra for *duck* onset



b. Difference spectra for *duck* onset



c. Transition spectrum 152-160 msec.

Figure 6.4 Static and transition spectra for onset demisyllable of *duck*

Steps in spectral analysis are shown above. The first step is conversion from Hertz to Bark frequency and Gaussian smoothing. The peak between 11 and 12 bark is the 2nd formant, dropping in frequency between the two sample times of 144 and 160 msec. The difference spectrum effectively normalizes the spectrum for changes in amplitude. The 2nd formant frequency shift is also evident. The 2nd formant's change is captured in the transition spectrum during the period of maximum spectral change. The positive value at 11 bark reflects the formant's drop in frequency and the increased amplitude of the difference spectrum at that frequency. The negative value just below 12 bark occurs behind the formant's motion, where the difference spectrum has fallen in amplitude.

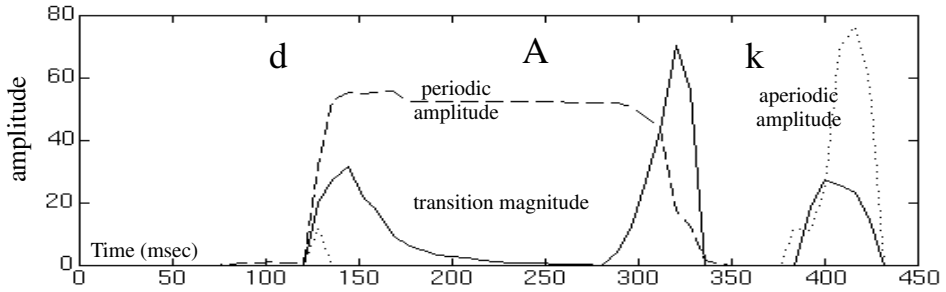


Figure 6.5 Transition magnitude for duration of *duck*

The transition magnitude τ (solid line) and amplitudes of periodic and aperiodic sounds (rescaled) are shown for the entire utterance in the graph above. Also shown are the approximate locations of linguistic segments. Spectra pictured in Figure 6.5 were sampled from the first transition peak between /d/ and /A/.

- **Transition magnitude:** To detect and distinguish stable and transient segments, the model computes transition magnitude, a measure of global spectral change which factors out the contribution of amplitude change. Transition magnitude τ is the L_1 norm of transition spectrum minus the L_1 norm of the difference spectrum.

$$\tau(t) = \sum_i |T_i(t)| - \sum_i |D_t(t)| \tag{6.1}$$

  The first term captures temporal changes in both total amplitude and spectral shape, the second captures only changes in amplitude. The result is averaged over three 8-msec frames. Figure 6.5 graphs the transition magnitude over the course of *duck*, together with periodic and aperiodic amplitudes (rescaled 50% and 30%, respectively).

- **Extreme transition detection:** A stable segment is a period during which the transition magnitude is at a local minimum or has remained below some threshold for more than some length of time since the previous segment. A transient segment is a period during which the transition magnitude is at a local maximum. These spectral stability and transience cues are combined with other cues to determine the segment type**.**

- **Sampled static and transition spectra:** Transition spectra are sampled for purposes of categorization only when a transient segment is detected. To do so, the model averages transition spectra over the 50 msec period centered at the point of maximum transition. Static spectra are likewise sampled only when a stable segment is detected, averaging over the 50 msec period centered at the point of minimal transition.

6.3.3 Learning spectral features

HABLAR uses soft competitive learning (Nowlan 1991a,b) to categorize sampled static and dynamic spectra. Because they represent different acoustic dimensions, static and dynamic spectra are categorized separately. We describe the computational details for static spectra; the process is nearly identical for transition spectra. Soft competitive learning may be done incrementally or as a batch process. Since

perceptual and motor learning are phased in this implementation of HABLAR, we use a batch process to learn spectral features.

The means of *N* Gaussians are randomly initialized, several more than necessary to represent the static sound categories in the linguistic environment. Gaussians have 44 dimensions (to match the 44 coefficients in each static difference spectra). Synthetic adult stimuli (described in the next section) are then presented to the model. The difference spectrum $\mathbf{D}^k$ sampled from each static segment detected among adult stimuli constitutes a separate observation *k*. For each observation, the probability $p_j(\mathbf{D}^k)$ that Gaussian *j* generated the observed spectrum $\mathbf{D}^k$ is

$$p_j(\mathbf{D}^k) = \frac{1}{K\sigma_j} e^{-\left\| \mathbf{D}^k - \vec{\mu}_j \right\|^2 / (2\sigma_j^2)}, \tag{6.2}$$

where *K* is a normalization constant, $\vec{\mu}_j$ are the means of Gaussian *j*, and $\sigma^2{}_j\mathbf{I}$ is its covariance matrix (assuming a spherical Gaussian distribution). Then the *conditional* probability of Gaussian *j* having generated the spectrum (given that spectrum $\mathbf{D}^k$ has been observed) is computed.

$$p(j|\ (\mathbf{D})^k) = \frac{p_j(\mathbf{D}^k)}{\sum_i p_i(\mathbf{D}^k)} \tag{6.3}$$

Finally, mean $\vec{\mu}_j$ of each Gaussian *j* is updated in proportion to the likelihood that it generated each observation.

$$\vec{\mu}_j = \frac{\sum_k p(j|\mathbf{D}^k)\,\mathbf{D}^k}{\sum_k p(j|\mathbf{D}^k)} \tag{6.4}$$

This process is repeated until learning converges. We then compare all resulting prototype patterns, deleting duplicates and prototypes whose activation is always below some threshold. This only makes the classification easier to interpret; it does not affect the nature of the solution.

6.4 Auditory Perception Simulations

Here we report results showing the properties of the transition and static spectra, the distinctiveness and faithfulness of its demisyllabic encoding. The results suggest that the segmentation algorithm is reliable and robust.

6.4.1 Stimuli

Stimuli are synthetic consonant-vowel (CV) syllables. Each stimulus employs one of three stop consonants (b, d, g) and one of ten vowels. A script language specifies a queue of gestures and proprioceptive triggering conditions necessary to produce each syllable type. Optimal vocal tract configurations necessary to accurately render each target vowel are determined, and gesture script templates are designed and fine-tuned by the experimenter for bV, dV and gV frames. Selecting target vowel and consonant at random, we generate 1,350 syllable tokens — about 45 tokens per syllable type.

To generate each token, Gaussian noise is added to each vowel's articulatory parameters. There is no way to determine if each resulting sound actually corresponds to the intended syllable type or even whether it is a legal sound of English without actually listening to it. A human listener evaluates each token, rejecting any non-English sound and rejecting any sound whose phonetic transcription does not agree with the intended type. The 892 remaining tokens are divided into training, validation, and test sets for supervised classification tasks.

6.4.2 Relative phonetic distance as a measure of distinctiveness

Each stimulus is presented to the auditory perception module, which segments it into prevoiced, transition, static, and silent segment types, sampling and storing transition and static spectra for later analysis. Phonetic distance Φ is measured among all syllable stimuli as a function of the linear correlation ρ between each pair of token transition spectra $\Phi(\mathbf{T}^i, \mathbf{T}^j) = 1.0 - \rho(\mathbf{T}^i, \mathbf{T}^j)$ (e.g., Pomerleau, 1993). We use this distance measure here because it factors out irrelevant differences in scale between two spectra. Other measures give comparable results.

The average phonetic distance between pairs of syllables of the same type is 0.211. The average distance between pairs of syllables of different types is 0.955. Thus, tokens of the same type appear to be relatively clustered in the representational space. A tabulation of average cross-distances for all bV syllable types appears in Table 6.2.

Table 6.2 Mean phonetic distance measured between bV syllable tokens

	bA	ba	bo	bO	bu	bU	b&	bE	bi	bI
bA	0.18	0.89	0.92	0.95	1.06	0.53	1.19	0.80	1.25	1.09
ba		0.33	0.88	0.72	0.94	0.92	0.81	0.96	1.15	0.93
bo			0.39	0.85	1.05	0.93	0.79	0.75	1.02	0.86
bO				0.13	0.96	1.03	0.86	0.91	1.29	1.12
bu					0.62	1.18	0.99	1.01	0.81	0.77
bU						0.33	1.23	0.79	0.21	1.21
b&							0.08	0.65	0.98	0.85
bE								0.13	1.06	0.63
bi									0.12	0.80
bI										0.09

Syllable types with the same vowel type but with different consonants (gE, dE, bE) are less distant with respect to each other (see Table 6.3) than syllables with the same consonant but different vowels (bE, bi, bo, b&, etc., below). Several tokens in the simulation corpus (especially Co and CO syllables) are not phonetically distinct. These tokens are also difficult for the experimenter to transcribe phonetically.

Table 6.3 Additional phonetic distance comparisons

Same Vowel

	bE	dE	gE
bE	0.13	0.55	0.50
dE		0.08	0.57
gE			0.10

Hard Discrimination

	bo	do	go
bo	0.39	0.52	0.39
do		0.43	0.39
go			0.27

6.4.3 Supervised classification as a test of faithfulness

As a test of the encoding's faithfulness, a multi-layer back propagation network is trained to classify syllables by eight features for vowel quality and consonantal place-of-articulation using transition spectra as input. The same test is performed with both transition and difference spectra as input, since some acoustic information is lost by the transition spectrum. As a control, we train a time-delay neural network (TDNN; Waibel et al., 1989) on the same task, using as input smoothed bark spectra over the entire duration of the consonant-vowel transition (averaging 168 msec). Each network is trained until a validation dataset indicates overtraining and then is tested using separate test dataset.

Supervised classification tests, which appear in Table 6.4, suggest that the transition spectrum is a robust demisyllable representation. Syllable classification error rate by the network which uses only the transition spectral coefficients as inputs is 5.29%. Error rates are substantially lower when both transition and difference spectra coefficients are used as inputs, dropping to 0.85% when both transition and difference spectra coefficients are rescaled to range between 0 and 1 for each pattern.[1]

Table 6.4 Supervised training error rates

Representation and Classification Method	**% Error**
HABLAR's Simple Transition Spectrum Representation (44 transition spectrum coefficients as inputs, tested with simple back-propagation classifier)	5.29%
HABLAR's Augmented Representation (44 transition and 44 difference spectra coefficients rescaled, tested with simple backpropagation classifier)	0.85%
Bark Spectrum Whole Syllable Representation (44 Bark spectrum coefficients for each of 21-8 msec frames, 3-frame input window, tested with TDNN classifier)	1.59%

The TDNN network's performance is similar. This is not surprising, as it must discover those static and dynamic features which are necessary to classify each syllable (Waibel et al., 1989).

Typical errors for all methods include single mistaken features, features just under threshold, or confusion in deciding the place-of-articulation.

Supervised classification results are encouraging for the use of transition data for segmentation and syllable classification without needing to learn segmentation strategies first. However, results point out that both dynamic and static spectral information is essential for accurate classification.

6.4.4 Locality and faithfulness of categorization

The model learns to accurately categorize static and dynamic sounds. To begin training, Gaussians are randomly initialized. Static spectra stimuli are then presented repeatedly, and each Gaussian is updated by changing its mean and covariance matrix in proportion to the likelihood that it generated the observa-

[1] The range of difference spectra coefficient values is much greater (by a factor of 10 to 20) than the range of transition spectra coefficients. Rescaling each set of inputs to the same range better conditions the weight space and leads to better results.

tion (Nowlan 1991a,b, see Section 6.3.3). Because this is an unsupervised learning algorithm, it is not possible to directly measure categorization error. Instead, we construct a confusion matrix.

Table 6.5 presents the results of static spectra categorization. Each column represents stimulus vowel types as classified by a human observer. Each row corresponds to a different Gaussian, whose parameters are initialized randomly but change during training such that each Gaussian tends to become most active when exposed to a particular class or subclass of sounds. Each is labeled by hand (for interpretability only) according to the vowel for which it is most active. Table entries in each row are the Gaussian's average activation when presented with a syllable containing the indicated vowel type. For example, among all syllabic stimuli containing [A], Gaussian #0 has an average activation of 0.69, Gaussian #6 has an average activation of 0.19, and Gaussian #7 has an average activation of 0.11.

Table 6.5 Unsupervised categorization of static spectra

##	Vow	[A]	[a]	[o]	[O]	[u]	[U]	[&]	[E]	[i]	[I]
0	[A]	**0.69**	----	0.01	----	----	0.05	----	----	----	----
1	[a]	0.01	**0.82**	0.05	0.02	----	0.02	----	----	----	----
2	[o]	0.01	0.07	**0.55**	0.11	----	0.02	----	----	----	----
3	[o]	----	0.05	**0.32**	**0.29**	----	0.01	----	----	----	----
4	[O]	----	0.02	0.04	**0.57**	----	0.01	----	----	----	----
5	[u]	----	----	----	----	**0.95**	0.03	----	----	----	----
6	[U]	0.19	0.02	0.02	0.01	0.02	**0.40**	----	----	----	----
7	[U]	0.11	0.02	0.02	0.01	0.03	**0.47**	----	----	----	----
8	[&]	----	----	----	----	----	----	**0.99**	----	----	----
9	[E]	----	----	----	----	----	----	----	**0.97**	----	----
10	[i]	----	----	----	----	----	----	----	----	**1.00**	0.01
11	[I]	----	----	----	----	----	----	----	0.02	----	**0.99**

The confusion matrix shows that the implicit categorization learned by the set of Gaussians is relatively localist — typically one Gaussian responds strongly to one vowel type. It also suggests that the resulting categorization is a faithful representation of vowel type. Similar results are obtained for the categorization of transition spectra.

This compilation of average activations confounds distributed activation patterns with outright categorization errors; hence, the representation's true faithfulness is not known. For example, Gaussians #6 and #7 seem to accurately classify [U], but on average they are also somewhat active when exposed to [A]. These average results could imply either consistently low activation Gaussian #6 and #7 for [A] tokens, or high activation and misclassification for a small percentage of [A] tokens. Of course, reality falls between these extremes. Some activation patterns are clearly unfaithful, but an inspection of activation patterns for all tokens reveals a generally faithful but distributed pattern of activation. The faithfulness of Gaussian activation patterns may be more definitively tested by using them as inputs for a supervised classifier. We have not performed such a test.

6.4.5 Emulating the acquisition of categorical perception

These results demonstrate a self-organizing process which learns a set of vowel and consonant-vowel categories based exclusively on acoustic cues and their statistical distribution in the linguistic environment. They appear to emulate children's phoneme-like discrimination of vowels (Kuhl et al. 1992) and consonants (e.g., Werker et al. 1981). Similar sounds are assimilated into existing categories, activating most that category (Gaussian) which is least distant acoustically. Dissimilar sounds activate distinct categories (Gaussians). Thus, the mixture of Gaussians simulates the "perceptual magnet" effect observed for vowel discrimination among all but young infants (Grieser & Kuhl 1989).

The model simulates only the "phonemic" perception of late infancy (see Section 2.8.2). It does not simulate its development except in the restricted sense of learning to categorize attended acoustic segments. Jusczyk (1993) has outlined a more complete account. Like ours, his WRAPSA model assumes low-level feature extraction, avoids traditional phonetic segmentation, and proposes a gestalt syllabic representation. To explain the development of selective phonetic listening, it proposes feature weighting or attention. Functionally similar behavior emerges from HABLAR's Gaussian mixture model. WRAPSA also integrates its perceptual machinery with a model of lexical access, something beyond HABLAR's scope.

Chapter 7 Hierarchical Motor Control

Having painted a picture of HABLAR's peripheral components, we proceed with a description of the cognitive architecture which links them. First we motivate the control architecture, identifying the functions it must serve and the empirical constraints it must satisfy. We then present heuristic and more formal pictures of the architecture and training methods, identify and resolve some additional computational issues, and finally detail the control architecture's implementation.

7.1 Empirical constraints governing motor control

At one end of the model, we have postulated an auditory perception module which generates a categorical, syllabic representation of speech sounds. At the other end, we have postulated articulatory control of the vocal tract and respiratory system which is organized into discrete gestures. The problem remaining to be solved is the mapping between the internal representation of the target utterance and its realization as a sequence of articulatory gestures. One line of evidence discussed in Chapter 2 suggests that the mapping is accomplished using some associative, stochastic mechanism. Other lines of evidence (Chapter 2 and Chapter 3) suggest a partition into problems of articulatory skill and linguistic composition. To determine where to partition the mapping, we examine how phonetic and articulatory events are related.

7.1.1 Phonetic vs. articulatory and linguistic events

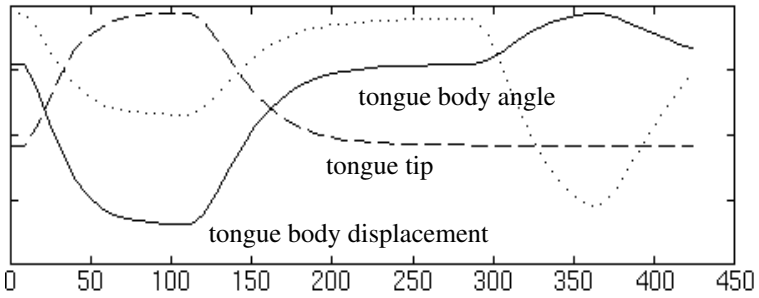
Auditory perception identifies several types of phonetic events — periods of silence, prevoicing, maximal spectral change, and minimal spectral change. These events and their properties are encoded as segmental and spectral features which constitute each utterance's phonetic representation. Each type of event has linguistic and articulatory significance. Linguistically, transitions usually signify demisyllables (consonant-vowel or vowel-consonant sequences), diphthongs, or glides. Static segments signify the stable portions of vowels, nasals, approximants, and fricatives. With respect to articulation, transitions correspond approximately to the midpoints of articulatory gestures. Static segments (including periods of silence and prevoicing) correspond to the ends of gestures. Moreover, distinctive nonlinear acoustic changes usually occur between a gesture's terminal configurations (Stevens, 1989). These enhance spectral transitions and ensure their reliability as linguistic codes and articulatory encodings. Figure 7.1 illustrates the relationship between transitions and gestures for a sample of synthetic adult speech. It graphically compares several key gestural components of *duck* (Figure 7.1a) with the measured transition magnitude (solid line in Figure 7.1b). Table 7.1 details the linguistic and articulatory significance of acoustic segments.

7.1.2 Coordinative structures and the partition of motor control

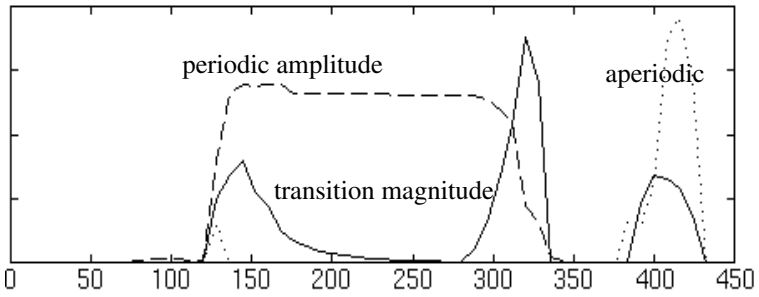
Upon closer examination, it becomes apparent that the crucial relationship between phonetic and articulatory events does not involve individual gestures but rather a set of functionally coordinated gestures. Even the very simplest of consonants involves a finely tuned coordination of vocalization and a ballistic gesture of tongue or lips to close or open the vocal tract. Though many sounds explicitly involve just a few articulators, they implicitly require the cooperation of several others. For example, /t/ and /d/ involve the tongue tip, but the lips must also be parted. This resembles what has been called a *coordinative structure*, a "temporary marshalling of many degrees of freedom into a task-specific functional unit" (Kelso et al. 1980, 1984, 1986; Saltzman & Munhall 1989). A coordinative structure achieves its functional end without regard to vocal tract perturbations, automatically compensating for external influences or articula-

Table 7.1 Acoustic segments vs. articulatory and linguistic events

Type	Articulatory basis	Linguistic relevance
Transition (T)	Zenith or center of gesture.	Demi-syllables, diphthongs
Static (S)	Beginning or end of gesture, when vocal tract is nearly stationary.	Static portion of vowels, nasals, approximants.
Static Friction (F), Voiced Friction (Z)	End of a gesture, during vocal tract constriction and air-flow with audible friction.	Static portion of fricatives.
Transition Friction (X)	Early in a gesture after the release of a constriction in the vocal tract but before voicing starts.	Aspiration after release of voiceless stop consonant. Some whispered sounds.
Burst (B)	Release of closure at start of gesture after lung pressure build up.	Burst after release of voiceless stop consonant.
Prevoicing (P)	Beginning or end of gesture as vocal tract is closed but vocal cord vibration continues.	Prevoicing of voiced stop consonant.
Silence (0)	Beginning or end of gesture as vocal tract is closed but without vocal cord vibration.	Silence of voiceless stop consonant closure (or silence of vocal tract at rest).



a. Several articulatory gesture trajectories during *duck*



b. Transition magnitude (solid), periodic (dashed) and aperiodic (dotted) amplitudes

Figure 7.1 Articulatory gestures vs. spectral transitions

tory contexts. The goal is to reproduce a minimal sequence of phonetic events which has linguistic significance.

This suggests a reasonable partition of motor control into two levels, phonological and articulatory, in which the articulatory control level's responsibility is the mastery of elemental, linguistically significant articulatory skills, and the phonological level's responsibility is mastery of linguistic composition. Articulatory control coordinates articulatory gestures to generate linguistically significant sounds without regard to vocal tract perturbations or articulatory contexts. In other words, the articulatory control level implements a set of coordinative structures, each a specialist in pronouncing some linguistically important sound. Phonological control composes linguistic events by choosing and activating sequences of articulatory specialists. The ability to correct for perturbations suggests a closed-loop control structure, at least for articulatory control.

7.2 Form and function of hierarchical motor control

HABLAR's motor control is carried out by a hierarchy of phonological and articulatory controllers. Each articulatory controller's responsibility is articulatory skill: the pronunciation of an elemental, linguistically significant sound. The phonological controller's responsibility is sound composition: combining elemental linguistically significant sounds.

Each controller performs what is called a *closed loop policy*, a plan which associates an action with each observation it makes of the system it controls. Each controller employs reinforcement learning to learn an optimal policy, a plan which maximizes cumulative rewards. Each is implemented as an artificial neural network whose inputs represent the controller's environmental state and whose outputs represent the utility of each possible action given the current environmental state. Each receives rewards related to the phonetic proximity of target and actual utterances.

This section presents the formal framework in which articulatory and phonological controllers are described, and presents each controller type heuristically and formally, detailing input, outputs, and computational roles.

7.2.1 Formalizing control problems as Markov decision tasks

Closed loop control problems faced by phonological and articulatory controllers may be formalized as Markov decision tasks. Doing so justifies the use of Q-learning and provides a framework with which to examine certain computational issues. A *Markov decision task* (first introduced in Section 3.5; see also Figure 3.1) is characterized by an *environment* to be controlled, the environment's *state*, a finite set of *actions* from which the controller chooses, *environmental dynamics*, and scalar feedback or *reinforcement*, the cost and/or payoff incurred by the agent as a function of its action in the current state. The environmental dynamics of a Markov decision task are formally expressed as a *transition function*, the set of probabilities of reaching one state from another given each action. A controller's *input* is its observation of the environmental state. Its *output* is its choice of an action. A closed-loop *policy* is a mapping from states to actions. An *optimal policy* is a state-action mapping which maximizes the objective function. The controller's task is often described in terms of a goal state to achieve or a trajectory of states to traverse. However, optimal policies are formally defined by a *schedule* of reinforcements and an *objective function* which computes a weighted or discounted sum of reinforcements over the course of each learning trial.

A Markov decision task must conform to *Markovian properties*, an additional set of constraints on environmental dynamics and rewards. One constraint requires that the current state be exclusively a function of the immediately past state, the immediately past action, and the unchanging set of transition probabilities which comprises the environmental dynamics. Likewise, reinforcement must be exclusively a

function of the immediately past state and immediately past action. Neither transitions between states nor rewards may depend on states or actions more than one time step in the past. Environmental dynamics must be *stationary*; that is, the set of transition probabilities must be fixed.

Q-learning is guaranteed to find an optimal policy for a finite Markov decision task which meets these properties and a number of additional formal constraints (Watkins & Dayan 1992).

7.2.2 Articulatory controllers

There are multiple articulatory controllers. Each is a specialist that generates all or some discrete portion of a familiar sound pattern. Given a target sound, the articulatory controller most likely to accurately reproduce it is chosen and activated by the phonological controller. Each articulatory specialist is a closed loop controller. It executes a closed-loop policy which associates the correct choice of articulatory gesture (its output) with each articulatory state it observes via proprioceptive feedback from the vocal tract and respiratory system (its inputs).

The articulatory controller is all but deaf. It knows only indirectly about sound, receiving positive reinforcement if the sound it produces is close to the phonetic target or some portion thereof. It is this distal phonetic reinforcement which ensures the linguistic significance of articulatory controller behavior. Between HABLAR's dual tasks of articulatory skill and sound composition, the articulatory controller addresses the former. It approximates the function of a coordinative structure whose goal is the generation of elemental linguistically significant sounds.

The time scale at which each articulatory controller operates is fine-grained and precise; it discovers and executes the exquisite gestural coordination necessary to produce each sound. Its job is not to choose each point through which each articulator moves. It only chooses the target equilibrium points of gestures in each articulatory dimension. It does not execute the gestures; it only selects and coordinates them.

Formalized in the context of a Markov decision task, the articulatory controller's observed environmental state is the set of proprioceptive and tactile sensations recognized by proprioceptive perception (see Chapter 5). These are summarized in Table 7.2. The same environmental state is available to all articulatory controllers.

The articulatory controller's actions are its choices of gestural equilibrium positions for each articulatory dimension. The number and range of gestural targets in each articulatory dimension are summarized in Table 7.3. For example, there are four lung pressure targets ranging from 0 to 10 cm H_2O and ten jaw angle targets between –12.6 and –18.6 degrees from horizontal as measured from the jaw fulcrum. Gestural details change slightly among simulations as we experiment with various articulatory constraints.

The articulatory controller's environment consists of gestural controllers, the articulatory synthesizer, and the source model. The environment's dynamical properties are a function of each gesture's damped-spring-like dynamics, the respiratory system's dynamics, and the interface between articulatory and gesture controllers.

The active articulatory controller observes a small constant cost for each discrete time step. It observes a reward only at the end of the period during which it is active, and only to the extent that the phonetic representation of the acoustic feedback approximates the phonetic representation of the target sound during its activation.

At each discrete time step (8 msec of real world time), gesture controllers update positions of all articulators and the source model computes air flow and changes in the respiratory state. The active articulatory controller observes the new environmental state and chooses an action. Affecting this process is a refractory period during which gestural controllers do not respond to new articulatory actions but continue

Table 7.2 Proprioceptive state and articulatory controller inputs

Class of inputs	Units	Quantities measured
Vocal tract configuration	13	Current position of jaw, tongue, lips, velum, etc.
Target configuration	13	Target position of vocal tract and lungs
Gestural phase	66	Measures progress of each articulator through current gesture (6 units per articulatory dimension)
Global tactile state	5	Closure, intensities or presence of vibration, turbulence
Tactile events	10	Trace history of contact at each place-of-articulation
Airflow events	5	Vibration, frication, aspiration
Respiratory state	6	glottal, oral, nasal airflow; tidal volume; subglottal and oral air pressure

Table 7.3 Summary of gestural targets

Articulatory dimension	Targets	Range of targets	Measurement
Extra-pulmonary lung pressure	4	0 to 10 cm H_20	inspiratory pressure
Glottal opening	7	–0.05 to 0.80 cm	width of opening
Jaw angle	10	–12.6 to –18.6 deg	from jaw fulcrum (from horizontal)
Tongue body distance	9	7.39 to 8.21 cm	from jaw fulcrum
Tongue body angle	13	–2.9 to –19.2 deg	relative to jaw angle
Tongue tip length	5	2.68 to 3.13 cm	from point on tongue body
Tongue tip angle	6	0.0 to 25.8 deg	relative to tongue body
Lip protrusion	6	0.76 to 1.69 cm	from upper incisors
Lip height	4	0.10 to 0.39 cm	closure relative to incisors
Hyoid anterior distance	4	7.14 to 7.37 cm	from fixed point
Hyoid superior distance	5	6.92 to 7.50 cm	from fixed point
Velar opening (na)	4	0.0 to 1.0	degree of opening

to maintain their old trajectories. The probability p_r that a new trajectory replaces an older one increases quadratically with the phase ψ of the old gesture according to the formula

$$p_r = \begin{cases} \min(1.0,\ (\psi - 1)^2 / k_\psi), \psi > 0 \\ 1.0, \psi = 0 \end{cases}. \tag{7.1}$$

Its purpose is to simulate hysteresis of recently activated gestural targets and to constrain gestural dynamics, reducing the random "jitter" of the untrained controller.

7.2.3 Phonological controller

The phonological controller's job is to choose which articulatory controller is best suited to generate the target utterance or some portion thereof. Like its articulatory subordinates, it is a closed loop controller; its actions are based on the phonetic feedback it receives. It executes a closed-loop policy which associates the correct choice of articulatory controller with each phonetic segment it observes in auditory feedback. Its goal is to generate a sequence of sounds whose phonetic representation comes to exactly match the target sound's phonetic representation. Between HABLAR's dual tasks of articulatory skill and linguistic composition, the phonological controller addresses the later. The time scale in which it operates is coarse-grained and event-driven.

The phonological controller's job may be formalized as a Markov decision task. The environmental state it observes has two parts, the phonetic representation of the target utterance, and the phonetic representation of the current utterance's acoustic feedback. Its action is the choice of 1-of-n articulatory controllers to activate. The phonological controller observes a small constant cost for each action it takes. It observes a reward proportional to the phonetic proximity of actual and target utterances. A trial ends when the phonetic feedback matches (or nearly matches) the target utterance or after a time-out period.

7.3 Relationships between control levels

The relationship between phonological and articulatory control levels is illustrated in Figure 7.2. Labor is divided by task. The phonological controller has a supervisory role relative to articulatory controllers. Environments of each control level are distinct, and control levels operate at different temporal granularities. The superior phonological controller negotiates a sequence of abstract states, but it need not decompose its environment or discover a set of abstract states on its own (see Section 3.5.2). Instead, abstract states are delineated independently by HABLAR's auditory perception. The phonological controller's environment encompasses the articulatory controller's environment. However, the phonological controller is ignorant of any articulatory details. It observes only the discrete phonetic representation of its auditory feedback.

7.3.1 Role of phonological controller in linguistic composition

The phonological controller plays several roles, some not immediately apparent, but all are necessary to ensure its ability to compose utterances out of elemental sounds. Its most apparent role is to map sequences of phonetic events (encoded in the representation of a target utterance) into sequences of articulatory events (the activation of articulatory controllers). Once the model's articulatory controllers have learned to generate a repertoire of elemental sounds which match phonetic events in the linguistic environment, then the phonological controller may learn to compose complex utterances out of elemental sounds,

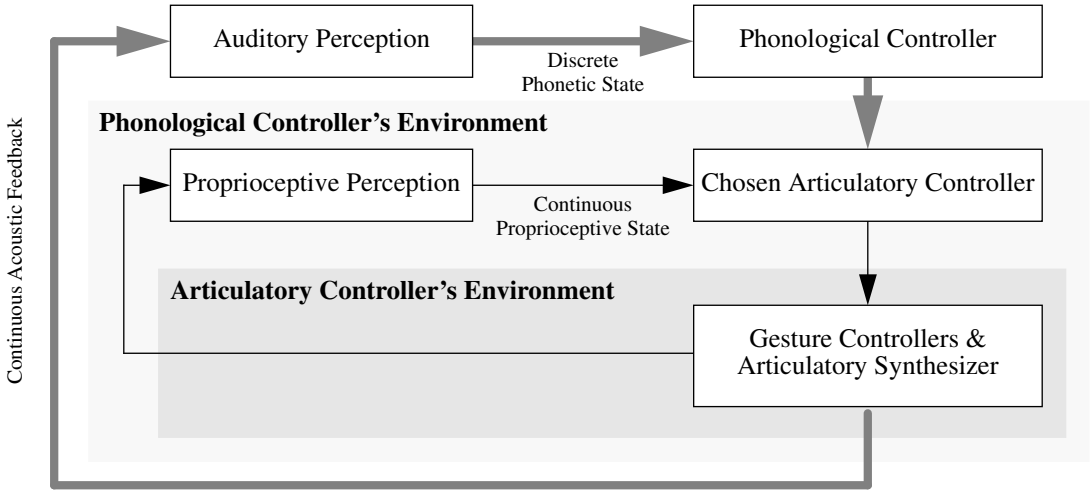


Figure 7.2 Nested phonological and articulatory control levels

learning the compositional structure of the language's phonology. Compositional properties implicit in the phonetic representation and the phonological controller's generalization properties contribute to accomplishing this goal.

What ensures that articulatory controllers generate linguistically significant articulatory events? Articulatory controllers receive rewards proportional to the phonetic proximity of target sounds and actually produced sounds. Target utterances — whether imitated or drawn from the child's lexical memory — reflect the linguistic environment in which the model is situated. Hence, sounds produced should come to resemble targets. What ensures that the *statistical distribution* of sounds generated by articulatory controllers matches the distribution of sounds in the linguistic environment? The phonological controller builds the equivalent of a statistical mixture model of the distribution of phonetic events in the linguistic environment. This statistical model is also the basis with which the phonological controller selects articulatory controllers to generate target sounds. All controllers receive rewards to the extent that this statistical mixture model mimics the distribution of sounds in the linguistic environment. We discuss this in greater detail in Chapter 9.

7.3.2 Role of the articulatory controller in linguistic composition

The ability to compose complex sounds out of simpler ones depends on the computational properties of both phonological and articulatory controllers. Compositional properties of the phonetic representation and other constraints discussed in Section 7.3.1 satisfies some requirements. What makes compositionality possible at the articulatory level is the seamless integration of control as responsibility for motor control is transferred from one articulatory controller to the next.

Two conditions are necessary for this to occur. (1) All articulatory controllers must share the same proprioceptive state, a condition met by HABLAR's design. (2) Instances of phonetic segment types must have similar proprioceptive states, even in different contexts, and articulatory controllers must ignore certain irrelevant proprioceptive events.

For example, consider what is necessary to generate *mad* [m&d] out of previously learned articulatory patterns for *bad* [b&d] and *man* [m&n]. The proprioceptive states corresponding to articulation of the vowel [&] in *mad*, *bad*, and *man* are all similar — the jaw nearly at rest after having opened for the vowel, tongue body likewise at rest in a position corresponding to the vowel, and vocal cords vibrating. The only condition which varies is the prior state of the velum (closed for /b/ and open for /m/). The articulatory controller's generalization properties must ensure that this difference is ignored.

What seems necessary for such composition to occur are articulatory controllers which evolve during training to approximate pure vowels and pure consonants. This may emerge partly out of the generalization properties of the articulatory controllers. It may be encouraged by constraints or rewards which allow the phonological controller to switch between articulatory controllers only during static acoustic segments including silence and prevoicing. Limiting the number of articulatory controllers may also encourage this.

7.3.3 Ideal operation illustrated

Figure 7.3 demonstrates the intended behavior of HABLAR, again using *duck* as an example. Time flows left to right and top to bottom. Each column summarizes the events associated with the activation of one articulatory controller. The first row of panels displays how the phonetic representation (shown as segmental and spectral feature types) changes over the course of the utterance and indicates the phonetic event which triggers the phonological controller's choice of articulatory controller. The second row of panels summarizes the articulatory gestures executed by the articulatory controller while it is active. Each hypothetical articulatory controller is labelled, but for interpretative purposes only. The third row of panels displays graphically the resulting changes in the vocal tract. The final row displays the cumulative acoustic consequences of the vocal tract's activity. Auditory perception in turn phonetically interprets this auditory feedback and updates its representation, as shown in the top panel of the next column.

Suppose the target sound has been chosen. Subsequent events may be summarized as follows:

- Initially there is silence. Given the phonetic target, the phonological controller chooses an articulatory controller that, when activated, closes the jaw, raises the tongue body and tip, and approximates vocal cords. Glottal vibration occurs during the alveolar closure, which results in prevoicing. This activity corresponds to the onset of the /d/.

- Prevoicing is detected by auditory perception, and the phonological controller chooses an articulatory controller that, when activated, observes the vocal tract closed at the alveolar ridge and then lowers the tongue tip and body while voicing continues. This produces the vowel /A/ and the transition between the onset /d/ and vowel.

- Auditory perception detects transition and static segments corresponding to [dA] and [A]. The phonological controller chooses the next articulatory controller, which raises the tongue body to the velum, starts to close the glottis, and adjusts lung pressure. This produces the /k/ and the spectral transition between /A/ and /k/.

- Transition segment corresponding to [Ag] and silence are detected, and the phonological controller chooses another articulatory controller. The glottis is opened, the lungs are relaxed, and tongue contact is released. A segment of dynamic friction corresponding to the aspirated velar release /k^h/ is detected, and the target phonetic representation is achieved.

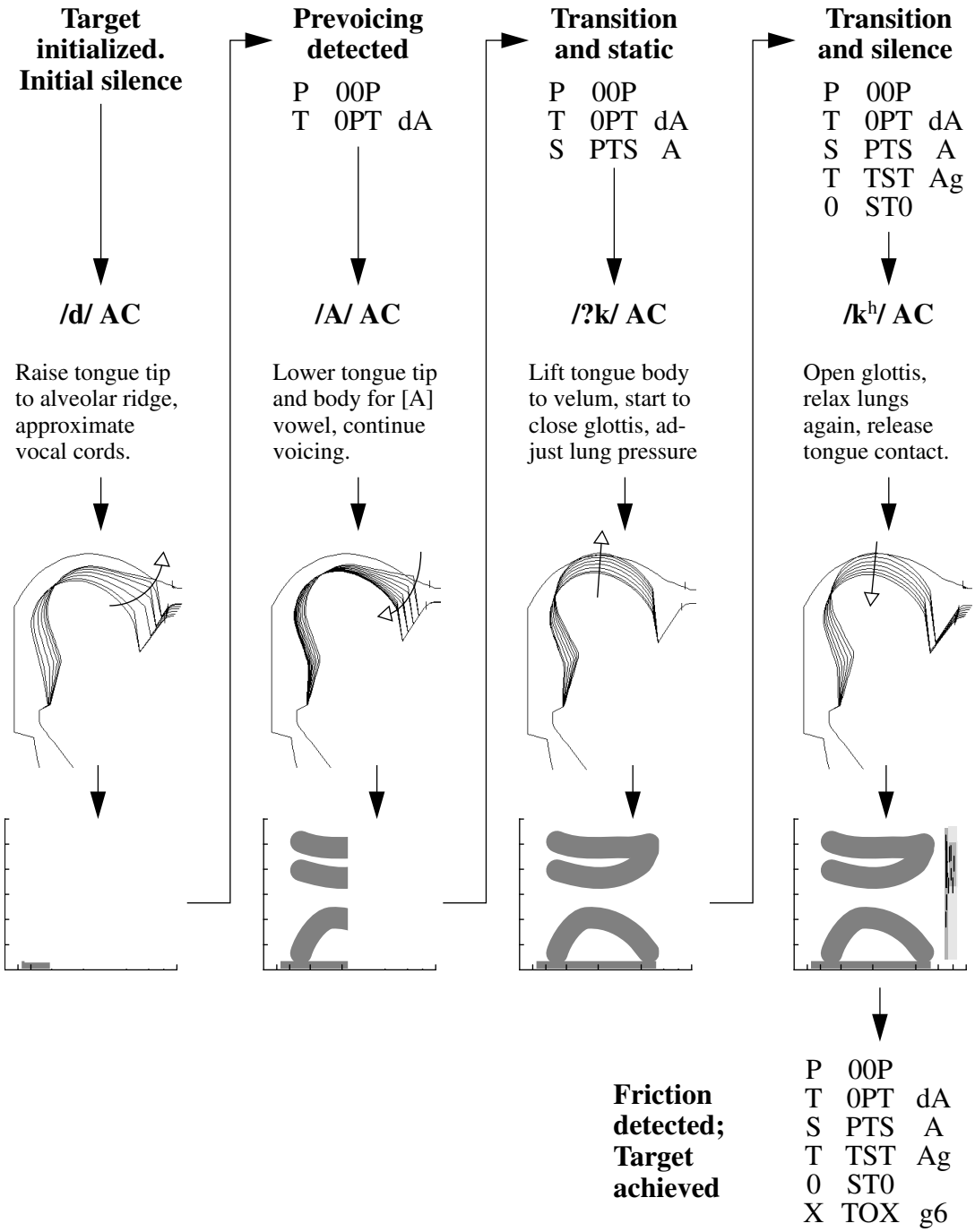


Figure 7.3 Intended HABLAR behavior

This figure shows the ideal sequence of events for HABLAR to pronounce *duck*. The top row of panels show how the phonetic representation changes over the utterance (including segment, triad, and spectral types as they are detected), the second row summarizes actions of the active articulatory controller, the third row portrays changes in the vocal tract, and the bottom panel shows the cumulative spectrogram of the resulting sound. Time flows from left to right and top to bottom.

7.3.4 How many articulatory controllers?

As a worst-case estimate, the number of articulatory controllers needed for English is equal to the number of demisyllables, diphthongs, and stand-alone static sounds in the language. This is some multiple of the number of spectral categories (to account for consonant voicing variations) estimated in Section 6.3.1. However, as apparent from the example presented in Figure 7.3 and discussion in Section 7.3.2, the best-case estimate is closer to the number of phonemes in English, possibly multiplied by a factor for anticipatory coarticulation. However, a full accounting of coarticulatory linguistic phenomena may require model refinements which could further reduce the required number (see Section 9.3.1).

7.4 Computational issues of control

Having formalized each control level as a Markov decision task, we are justified in using Q-learning to the extent that formal constraints are satisfied or approximated. However, several appear to be violated. State spaces at each control level are continuous, not finite and discrete. Some states are hidden, and some rewards depend on past states and actions, violating Markovian properties. The hierarchical relationship of control levels may also violate Markovian properties, may perturb otherwise stationary environments, and may make search more complex. In the following sections, we explore these issues in greater detail and propose means to address them to enable effective learning of articulatory and phonological control. We start, however, by answering an obvious question postponed from earlier chapters.

7.4.1 Why reinforcement learning?

Unlike the articulatory skill models discussed in Chapter 3, HABLAR's motor control task is not defined by an acoustic or articulatory reference trajectory. Rather, the task is specified by the target phonetic representation, and task success is determined by comparing the target with the phonetic representation of the actual utterance.

Superficially, this resembles a supervised learning problem. A signed error vector may be computed from target and actual phonetic feature vectors at the end of the utterance. Why not use supervised learning methods to learn the system's forward dynamics and acoustics? If we do, then motor controllers must determine both system dynamics and the additional temporal credit assignment problems posed by delayed phonetic feedback. Learning the system's forward dynamics could be a formidable task. Fortunately, hard-wired gestural and pulmonary dynamics simplify the control task, obviating the need for a gestural dynamics forward model. Under these circumstances, Jordan and Rumelhart (1992) suggest using an integrated quantity which measures overall task success to train a forward model. By introducing a scalar integrated quantity, we effectively transform the task into a reinforcement learning problem (see Section 3.4.4).

If supervised methods are used on the reformulated problem instead of reinforcement methods, unnecessary computational costs may be incurred. This is most likely to occur at the phonological control level. Phonetic representations are relatively local and nearly discrete. The phonological controller learns an approximately discrete mapping from phonetic target and feedback representations to the discrete choice of articulatory controller using an integrated quantity or an evaluative signal as the training signal. This is analogous to a set of tasks used by Markey and Mozer (1992) to compare the performance of forward model based and stochastic gradient-following reinforcement learning methods. Training time for the forward model method is one or two orders of magnitude slower than for the fastest stochastic method. Further analysis shows that this is caused by the forward model's counterproductive generalizations. Although forward model performance is improved by introducing additional structure (decomposing it into

separate forward models for each subtask), the best stochastic reinforcement learning method still outperforms the forward model method by two-to-one (unpublished data).

Ultimately, the choice between supervised and reinforcement methods must be empirical, but a comparison is beyond the present scope of this research.

7.4.2 Continuous-valued state variables

The articulatory controller's proprioceptive domain and the phonological controller's phonetic domain are continuous, multi-dimensional spaces. Q-learning requires a finite, discrete state space. To overcome this restriction, some form of generalizing function approximation is typically used to accurately estimate the Q-function, simplifying and abstracting each controller's huge domain into a set of internal features which capture functional distinctions among input states. HABLAR employs this strategy.

7.4.3 Function approximation and aliasing

Using function approximation to estimate the Q-function comes at a cost. The environmental state and the "state" actually observed by the articulatory controller are not the same. Useful abstraction divides observed environmental states into classes which are as large as possible without introducing irrelevant distinctions and which are functionally equivalent with respect to the set of actions required for optimal control. *Perceptual aliasing* occurs when abstraction confounds environmental states otherwise necessary for optimal control (Whitehead 1992). Network generalization may also confound or obscure input states. This is a trade-off HABLAR cannot entirely avoid.

By judicious choice of network architecture, learning parameters, and input representations, generalization-based aliasing might be reduced. HABLAR borrows the architectural strategy employed by Lin (1992). One typical network Q-function approximator uses a single network with input units for state *and* action, sigmoidal or Gaussian hidden units, and a *single* linear output unit which represents the Q-value. Lin instead uses a separate output unit or an entirely separate network (Lin, personal communication) for each action's estimated Q-value; the network's input units represent only environmental state variables. This approach reduces or eliminates generalization among actions but does not prevent generalization among related environmental states. HABLAR adopts this approach.

7.4.4 Ensuring Markovian rewards

Markovian properties require that neither state transitions nor reinforcements depend on states and actions more than one time step in the past. State transitions satisfy these constraints at both control levels. However, without augmenting the articulatory controller's observed state with information about past states, its rewards do not satisfy Markovian properties. A whole trajectory of past states and actions, not the final action and state, are rewarded. This is because articulatory controller rewards are a function of phonetic feedback, which encodes acoustic events spanning an entire acoustic segment and thus an extended sequence of actions.

Consider the two idealized articulatory policies represented in Table 7.4 and Table 7.5. The goal is to pronounce /ba/. Reinforcement is determined only after the whole syllable is perceived. One policy succeeds; the other does not. An agent attempting to learn this task with an unaugmented state cannot learn a well-defined Q-function or closed-loop policy, since some equivalent state-action pairs accrue a reward, and some do not. The state labeled "voice ON, oral cavity OPEN, NO lip contact" and action labeled "OPEN glottis" leads to reward in Table 7.4's policy but not in Table 7.5's policy. Prior actions which close and open the jaw distinguish the successful policy. In the articulatory world which HABLAR simulates,

many more time steps (each 8 msec) and gestural targets intervene between actions and reward. The consequences of delayed phonetic feedback can be much worse.

Table 7.4 Idealized articulatory sequence which generates /ba/

State			Action
Voice	**Oral cavity**	**Lip contact**	
off	open	no	close jaw
off	closed	yes	exhale, approximate vocal folds
on	closed	yes	open jaw
on	open	no	open glottis
off	open	no	Reward = 1

Table 7.5 Articulatory sequence which does not generate /ba/

State			Action
Voice	**Oral cavity**	**Lip contact**	
off	open	no	exhale, approximate vocal folds
on	open	no	open glottis
off	open	no	Reward = 0

Incorporating memory of past states into the current environmental state avoids this problem. Lin & Mitchell (1992) investigate several possible neural network approaches to accomplish this, including memories formed by recurrent networks, but such general solutions are computationally expensive. A more economical solution relies on domain-specific knowledge and HABLAR's overall sensorimotor architecture. Articulatory controller memory need only be long enough for simple sounds, since the phonological controller composes longer ones, and its memory need only track categories of past actions which adequately distinguish categories of speech sounds.

HABLAR's proprioceptive perception therefore includes exponentially decaying memory traces of phonetically relevant tactile events (contact and release at locations corresponding to consonantal places-of-articulation) and recent occurrences of glottal vibration and vocal tract turbulent airflow. These are adequate for stop-consonant-vowel syllables, but a richer representation is necessary for more complex morphemes.

Even though its reinforcement is a function of phonetic feedback for the entire utterance, the phonological controller does not face the same problem because of its superpositional phonetic representation and contextual features. It faithfully captures the phonetic history of the entire utterance.

7.4.5 State isolation and reward hiding

Watkins (1989) explains that a simple-minded hierarchical arrangement of agents which directly observe and control the same state violates the Markovian constraint that state transitions be exclusively a function of current state and action. No violation arises as a result of articulatory and phonological controllers observing and controlling the same state. They control and observe different environmental states. The former observes and controls only the proprioceptive state. The latter observes and controls only the phonetic state. This conforms to the principle of state isolation proposed by Dayan and Hinton (1992) to avoid

situations like those identified by Watkins. Their principle of reward hiding is also enforced. The articulatory controller's reward is based on a temporally local phonetic representation, not the representation of the entire utterance.

7.4.6 Nonstationary search and hierarchical search complexity

The phonological controller's environment includes articulatory controllers whose expected behaviors change as they learn to pronounce various sounds. Thus, the phonological controller's environment is nonstationary, which may introduce Q-learning convergence difficulties. As a practical matter, this does not appear to be a problem. Articulatory controller behavior changes, but relatively slowly, and in conformity with the expectations of the phonological controller. This problem does not apparently affect the performance of other hierarchical reinforcement learning models (Chapter 3). Nonstationary articulatory specialist behavior is not unlike nonstationary expert behavior in a mixture-of-experts model (Jacobs et al. 1991) in the sense that gradient descent occurs on two levels simultaneously and learning on one level affects optimal behavior of the other level. What possibly sets HABLAR apart is its overall computational complexity, especially the search complexity represented by phonological and articulatory controllers together.

7.4.7 Bootstrapping the hierarchy

Hierarchical reinforcement learning is hard (Singh 1992c), and there do not appear to be any principled methods for reducing the difficulties (Singh, personal communication). To help overcome the combinatorial complexity faced by HABLAR's control hierarchy and to alleviate effects of the phonological controller's nonstationary environment, bootstrapping strategies are introduced which pace the complexity of target sounds. Phased complexity has several precedents, including Singh's (1992a-c) optional practice of training a hierarchical control system first on elemental tasks (see also Elman 1993, Tham & Prager 1994).

Three bootstrapping strategies are employed in simulations of the model: an engineered hierarchy, intentionally phased complexity, and emergent complexity.

Engineered hierarchy. The first strategy is intended more as a tool to initially test the feasibility of training articulatory controllers with phonetic targets and to test the feasibility of composing sounds with the phonological controller. Articulatory controllers are trained individually on component sounds, then the phonological controller is trained to compose a longer utterance using the learned capabilities of articulatory controllers, whose weights have been frozen.

Phased complexity. In the second strategy, HABLAR is presented with a set of parental sounds to be learned. Their complexity is intentionally staged. Articulatory controller temperature annealing rates are constant, but phonological controller temperature is a function of HABLAR's goal (high temperature for play and low for imitation) and its success (increased temperature after a series of failures). This is intended to mimic children's gradual loss of articulatory plasticity from birth until age 5 years but their continued compositional experimentation and word play.

Emergent complexity. This strategy, involving least direct experimenter intervention, is intended to replicate the pacing of target sound complexity which might emerge from a complete sensorimotor system whose several components pace each other.

- The lexicon learns sounds of increasing phonetic complexity (e.g., number and diversity of acoustic features) only gradually, pacing the complexity of target words. Due to the correspondence of

acoustic features, acoustic segments, and articulatory gestures, control of lexical complexity should also control phonological complexity.

- Lexical preferences based on semantic or pragmatic features or phonetic sound preferences are introduced to bias the earliest target sounds.

- Annealing strategies are the same as for phased complexity.

Motor and perceptual learning occur in distinct phases in HABLAR, a simplification of the real world. For human children, learning of phonetic categories (4-8 months for vowels and stop consonants) and babbling (6-12 months for canonical babbling) overlap. Phonetic categories learned earliest are likely to be more distinctive and more prevalent than those learned later. Thus, in a model in which perceptual and motor control learning overlap, the former could pace the latter. This is beyond the scope of this research at the present time.

7.5 Controlling HABLAR's multi-dimensional action space

Of the computational issues faced in the design of the articulatory controller, tractability is most serious. Consider a neural network approximator for a Q-function. If conforming to Lin's (1992) architecture (see Section 7.4.3), the network has one output for each possible action a. It takes as input a representation of the environmental state $\mathbf{x}$. Its outputs represent the value of the Q-function $Q(\mathbf{x}, a)$ for each possible action a given the current environmental state. In a multidimensional space, the number of possible actions can grow quite large, spanning the Cartesian product of the set of actions in each dimension. In standard Q-learning, there must be a separate output unit for each such action. This is illustrated for a hypothetical two-dimensional 3x4 action space in Figure 7.4, in which there are output units representing Q-values of all 12 combinations of actions.

Now consider the number of possible gestural target combinations from which the articulatory controller must choose. Two billion combinations are possible out of the twelve dimensions of articulatory gestures listed in Table 7.3. Nearly every configuration is anatomically possible, and it is impractical to

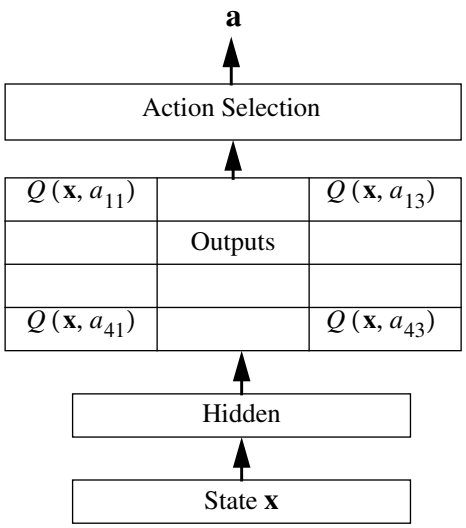


Figure 7.4 Standard Q-agent architecture

exhaustively evaluate all combinations. Any architecture which enumerates all possible action combinations faces similar problems.

To overcome this limitation, HABLAR decomposes the control task by degree-of-freedom. It employs a collection of cooperating, semi-autonomous subagents which operate in parallel, each subagent controlling a single articulatory dimension. Their collective behavior replaces the actions of a single agent. All subagents observe the same environmental state and reinforcement schedule, working together to maximize their common goal. Choice of action in each dimension is made independently of actions chosen in other dimensions. Though their actions are independent, subagents share a single network's hidden units in an attempt to cooperatively develop internal state representations which help solve the controller's overall task. We refer to this parallel Q-learning architecture as quasi-independent. Quasi-independent subagents make action choices by comparing Q-values within only a single articulatory dimension. The controller's action is composed from each component choice.

Figure 7.5 illustrates the network architecture for the same hypothetical two-dimensional 3x4 action space considered above. Instead of twelve output units, there are two groups. One group of four units represents the Q-values of four actions in one dimension; the other group of three units represents the Q-values of actions in the other dimension. Actions are chosen in the first dimension independently of the choice in the second dimension. The joint action is composed of these two choices.

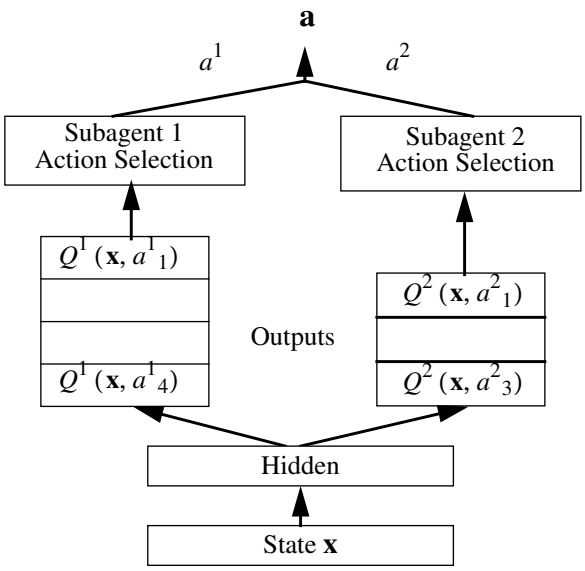


Figure 7.5 Quasi-independent subgent architecture

7.6 Controller implementation

We now present details of phonological and articulatory controller implementations, including reinforcement functions for each.

7.6.1 Neural network implementation of controllers

Each controller is implemented with an artificial neural network whose *inputs* take a representation of the environmental state **x**, and whose *outputs* represent the Q-values $Q(\mathbf{x}, a)$ of each possible action a, given the current environmental state **x**. Actions with the highest Q-value are the most probable and are chosen accordingly. Optimally, an agent chooses the most highly valued action, but it experiments with possibly suboptimal actions until an optimal plan is learned. $Q(\mathbf{x}, a)$ 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 thereafter.

7.6.2 Reinforcement functions

Phonological controller reward R is zero except at the end of each utterance. At that time, a reward signal is generated that is related to the distance between phonetic representations of target utterance and actual feedback. HABLAR uses squared Euclidean distance to measure phonetic distance. If t_i and f_i are activations of each phonetic feature i in target and feedback representations respectively, then the phonological controller's reward is

$$R = K_R \max\left(0, \sum_i t_i^2 - \sum_i (t_i - f_i)^2\right) \tag{7.2}$$

where K_R is a constant. When distance is zero, and the actual utterance exactly matches the target, the reward is proportional to the length of the target utterance as measured by the number of phonetic features in the target, i.e., $\sum t_i^2$. For distances other than zero, reward is reduced proportional to the square Euclidean distance. Reward is never less than zero, because phonetic distances greater than $\sum t_i^2$ do not seem to provide useful information about the error surface.

The active articulatory controller observes a reward computed in much the same way, but only for that portion of the utterance which occurs while it is active. This is done by maintaining a phonetic representation $\mathbf{f}^L$ of the acoustic feedback which is *local in time* to the active articulatory controller. At the beginning of the utterance or any time at which the active articulatory controller changes, activations of all features in this local representation are reset to zero. Thereafter, a representation is formed from the superposition of phonetic features detected while the articulatory controller is active. If f^L_i are activations of each phonetic feature i in the *local* feedback representation, and K_r is a constant, then the articulatory controller's reward r is computed at the end of the period during which it is active as follows:

$$r = K_r \max\left(0, \sum_i t_i^2 - \sum_i (t_i - f^L_i)^2\right). \tag{7.3}$$

Phonological and active articulatory controllers respectively incur small constant costs C and c for each action they undertake. The cost's cognitive justification is its representation of the effort expended to undertake each action. Computationally the cost is needed to build an implicit representation of environmental dynamics into each controller's Q-function, since there is no explicit representation of the transition function. Without some cost or discount of future rewards, all states on the trajectory to a goal would have the same value, and there would be no way to determine which way to move between two states.

7.6.3 Phonological controller implementation

The neural network implementing the phonological controller has sufficient inputs to accommodate the current phonetic feedback **f**, target phonetic representation **t**, and one output unit for each articulatory controller. Hidden units are sigmoidal; output units are linear. The target phonetic state **t** 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 feedback **f** is detected, which is a function of auditory perception. Its *i*-th output is the estimated utility $Q(\mathbf{t}, \mathbf{f}, a_i)$ 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 $p(a_i | \mathbf{t}, \mathbf{f})$ across all possible actions $a_i \in A$. Temperature *T* determines the randomness of action selection and is varied during learning according to an annealing schedule, the chosen goal strategy, and a bootstrapping strategy.

$$p(a_i | \mathbf{t}, \mathbf{f}) = e^{Q(\mathbf{t}, \mathbf{f}, a_i)/T} \Big/ \sum_{a_k \in A} e^{Q(\mathbf{t}, \mathbf{f}, a_k)/T} \tag{7.4}$$

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 when auditory perception detects a new phonetic segment and the phonological controller observes a new phonetic feedback $\mathbf{f}'$. At this time, the controller will also compute 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 the network's prediction of action *a*'s utility and use it to update the controller's parameters. The new predicted utility is the sum of cost *C*, reinforcement *R*, and the maximum utility among all actions in our new state $\mathbf{f}'$. The error is the difference of this sum and $Q(\mathbf{t}, \mathbf{f}, a)$, the utility of the chosen action.

$$E = R + C + \max_{a_k \in A} \{ Q(\mathbf{t}, \mathbf{f}', a_k) \} - Q(\mathbf{t}, \mathbf{f}, a) \tag{7.5}$$

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.

7.6.4 Articulatory controller implementation

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 $\mathbf{G} = G^1 \times \ldots \times G^n$. 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). Instead of learning a complete Q-function, each subagent explores only 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; subagents cooperatively explore the entire action space.

- There is one network whose input is proprioceptive state **s**. 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 $Q^j(\mathbf{s}, g^j)$ 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. That is, the probability of choosing action **g** is

$$p(\mathbf{g}) = \prod_{i \in [1 \ldots n]} p(g^i) . \tag{7.6}$$

- The Boltzmann distribution is computed only over actions in G^j and only over outputs in group j.

$$p(g^j_i | \mathbf{s}) = e^{Q(\mathbf{s}, g^j_i)/T} \Big/ \sum_{g^j_k \in G^j} e^{Q(\mathbf{s}, g^j_k)/T} \tag{7.7}$$

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

$$E^j = r + c + \max_{g^j_k \in G^j} \{ Q^j(\mathbf{s}', g^j_k) \} - Q^j(\mathbf{s}, g^j) \tag{7.8}$$

- *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.

The architecture is presented here as it applies to control of articulatory gestures. However, it is applicable to any n degree-of-freedom control problem in which selection of actions in each control dimension is independent per Equation (7.6).

This approach was introduced by Markey (1994a). A similar approach was developed independently by Tham and Prager (1993). Tan (1993) pioneered the use of cooperating Q-learning agents on a joint task.

7.7 Motor control simulations

Results of simulations demonstrating the stand-alone behavior of HABLAR's motor control and behavior of the integrated sensorimotor model are presented in Chapter 8.

7.8 Why does parallel Q-learning work?

Parallel Q-learning has now been applied to a number of problems, including a phonological task which resembles a 4-dimensional maze, control of a realistic vocal tract model (Markey 1994a, see also Chapter 8), more traditional 2-dimensional mazes and other problems designed to induce failure by the method (Markey in preparation), and a simulated robot with a 2-joint arm (Tham & Prager 1993, 1994). Why does the technique succeed without an explicit representation of the utility of all possible action combinations and without any explicit communication or planning among subagents? How might it fail?

7.8.1 Cooperation emerging from self-interested behavior

Consider the problem from the point of view of one subagent exploring its articulatory dimension. It attempts to maximize its future return without explicit regard for the behavior of others. A subagent's estimate of its *own* action's value is not a simple function of its own isolated behavior. Rather it is a function of the joint behavior of all subagents. Since the subagent is not privy to other subagents' "planned" behavior for the current time step, it can only estimate it. Its value estimate for its own actions in the current state is a function of the *expected behavior of the other subagents, conditioned on the current state*. A complete expression of subagent j's Q-value for action a^j in state $\mathbf{x}$ is thus:

$$Q^j(\mathbf{x}, a^j) = Q^j(\mathbf{x}, a^j \mid \{\mathbf{E}\{a^k \mid \mathbf{x}\}, k \neq j\}), \quad j = 1 \ldots n \tag{7.9}$$

All agents are linked through a common shared state. Hence, current state — together with the distribution of other subagents' actions as conditioned on current state — is the implicit mechanism to "transmit" information about what agents are likely to do.[1]

The method succeeds *not* because subagents make mutually independent decisions. As we have argued, the decision along one dimension is based on the expected subagent behavior in other dimensions. Instead, each subagent's estimate of the value of its own action must ultimately approximate the value of the joint action $\mathbf{a}$ of all subagents in state $\mathbf{x}$.

$$\max_{a^j \in A^j} [Q^j(\mathbf{x}, a^j \mid \{\mathbf{E}\{a^k \mid \mathbf{x}\}, k \neq j\})] \rightarrow \max_{\mathbf{a} \in A} [Q(\mathbf{x}, \mathbf{a})], \quad \forall j \tag{7.10}$$

This depends on cooperation emerging from the subagents' self-interested behavior (Barto 1985), which seems likely given their common state and reinforcement schedule. There are several empirical results for stochastic gradient-following reinforcement algorithms (terminology due to Williams 1992), which easily learn multi-dimensional vector actions (Ackley & Littman 1990, Gullapalli 1993, Markey & Mozer 1992, Prescott & Mayhew 1992, Williams & Peng 1989).

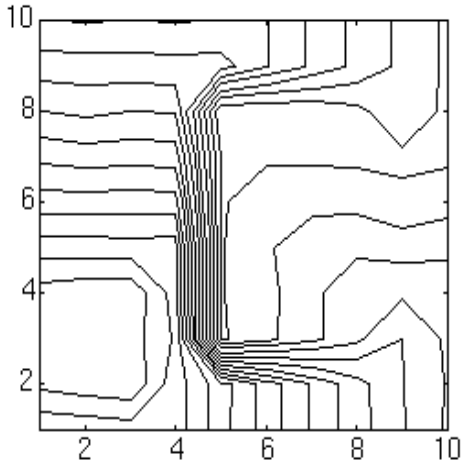
Equation (7.10) is equivalent to saying that each subagent learns the value function for the joint task. The *value function* measures the discounted return of following the optimal policy starting in the current state $\mathbf{x}$ and is related to the Q-function.

$$V(\mathbf{x}) = \max_{\mathbf{a} \in A} Q(\mathbf{x}, \mathbf{a}) \tag{7.11}$$

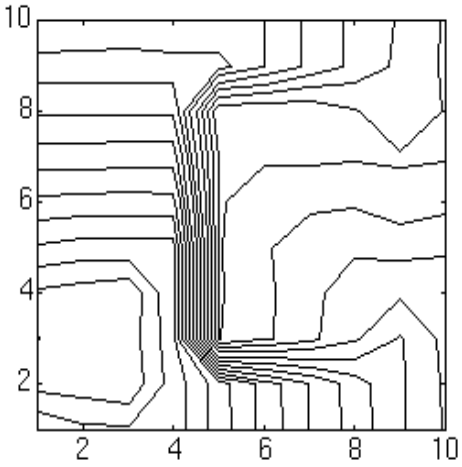
In the limit $V^j(\mathbf{x}) = V(\mathbf{x})$. Empirically, this is the case in all successful simulations.

One example for which we have simulation results involves learning to negotiate the 10 x 10 maze pictured in Figure 7.6 (lower left panel). An agent moves through the maze in any of 9 directions, N, NW, W, SW, S, SE, E, NE, and at-rest. The agent must negotiate around a wall, also pictured. Each trial starts with the agent placed at some random location in the maze and ends when the agent reaches the goal. One set of simulations employs the standard Q-learning architecture with a single agent with 9 possible actions (each of the above moves). Another set of simulations employs parallel Q-learning with two quasi-independent subagents. The action space is decomposed into two dimensions. The first subagent has 3 possible actions: N, S, and neither. The second subagent has 3 possible actions: E, W, and neither. The single agent

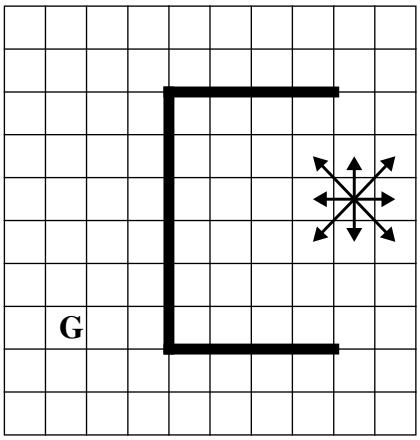
[1] I am indebted to Steve Nowlan for this suggestion.



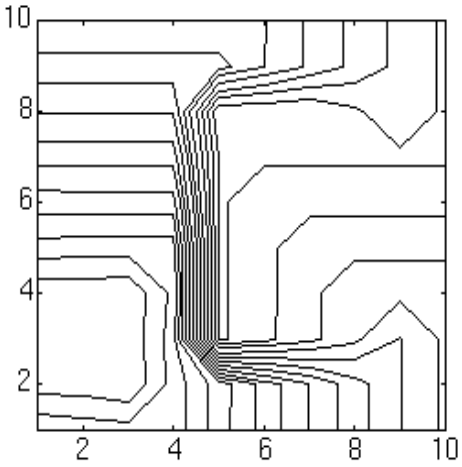
Subagent 1's value function



Subagent 2's value function



Maze



Standard Q-learnng agent's value function for the same maze

Figure 7.6 Value functions for standard and parallel Q-learning

Value functions generated by standard single-agent (lower right) and quasi-independent (upper left and right) architectures for a traditional 10x10 maze-exploration task (lower left).

learns the task in 8220 trials (632 standard deviation); the quasi-independent subagents learn the task in 3560 trials (743 sd). Besides solving the problem faster, the parallel Q subagents learn virtually identical value functions (top panels) which are almost indistinguishable from the value function learned by the standard Q-learning architecture (lower right panel).

These observation together with the empirical success of multi-dimensional control with gradient-following algorithms suggests using the Adaptive Heuristic Critic architecture (Barto et al. 1983) with one critic and many controllers (Dayan, personal communication). This has been implemented successfully by Tham and Prager (1993) for a two-dimensional robotic control problem.

7.8.2 Limitations of parallel-Q's sparse representation of action

The parallel-Q architecture, as applied to problems in which action is decomposed by degree-of-freedom, assumes that action choices in all dimensions are independent. Formally, that means that the probability of the joint action $\mathbf{a} = (a^1, \ldots, a^n)$ is the product of the probabilities of each subagent's action in each action space dimension (7.6).[1]

This implies that not all possible policies can be represented by a parallel-Q architecture. Only policies which admit to an outer-product decomposition may be represented (Robert Dodier, personal communication). However, we conjecture that all states are accessible by at least one optimal policy. An example is an XOR-like task in which *only one of two lights* must be turned on. The set of all possible policies is represented by the following matrix.

Probability of actions	Turn Light 1 OFF	Turn Light 1 ON
Turn Light 2 OFF	0.0	0.5
Turn Light 2 ON	0.5	0.0

This does not admit to an outer product representation, and parallel-Q may not represent both policies. However, it may represent either one or the other possible optimal policy, e.g., either

Probability of actions	Turn Light 1 OFF	Turn Light 1 ON
Turn Light 2 OFF	0.0	1.0
Turn Light 2 ON	0.0	0.0

or its transpose

Probability of actions	Turn Light 1 OFF	Turn Light 1 ON
Turn Light 2 OFF	0.0	0.0
Turn Light 2 ON	1.0	0.0

In simulations, 56% find one policy and 44% find the other. Similar instances occur in mazes in which there are two possible routes around an obstacle. Parallel Q simulations discover one of the two routes. This is an outcome of Q-learning's greedy behavior.

The 2-light problem is an example of a cooperative game. In general, cooperative one-step games do not necessarily converge to a global optimum. As formulated above, both local optima are also global optima because each light is equally rewarded per the following reward matrix:

---

[1] This discussion is due in part to conversations with Robert Dodier.

Symmetric reward contingencies	Light 1 OFF	Light 1 ON
Light 2 OFF	-1	1
Light 2 ON	1	-1

However, if the contingencies favor one light over another, we observe two Nash equilibria, stable states in which no agent benefits by unilaterally changing its strategy (Nash 1950, see Narendra & Thathachar 1989).

Asymmetric reward contingencies	Light 1 OFF	Light 1 ON
Light 2 OFF	-1	1
Light 2 ON	4	-1

The situation in which light 1 is ON with light 2 OFF is a local but not global optimum. In one-step games, this is likely to lead to suboptimal solutions. However, parallel Q-learning is not a one-step game but a dynamical system in which each agent gets to observe the results of its actions and correct its policies. Simulations given the asymmetric reward contingencies find the optimal policy 97% of the time. We are still attempting to understand why and when parallel Q-learning solves cooperative tasks despite Nash equilibria.[1]

7.8.3 Nonstationary environments and parallel subagents

Parallel-Q decomposition of a task leaves the state and reward structure unaffected. The transition of states observed by each subagent is still Markovian (if the undecomposed environment is Markovian). The environmental dynamics of the joint task are the same as the environmental dynamics of the undecomposed task. However, no subagent can know the transition function for the joint task. The environmental dynamics it observes are limited to the effect that its own actions have on state transitions, given the estimated behavior of other subagents. Since each subagent's behavior changes, individually for each subagent the environmental dynamics appear to be nonstationary. This issue does not appear to be of practical concern, since all subagents coevolve to achieve the same goal. However, it does violate the formal constraints under which Q-learning is currently known to succeed.

7.8.4 Empirical success

A formal analysis of task decomposition by degree-of-freedom is needed, but is beyond the scope of this dissertation. Empirically it appears to be robust, limited only by unrelated issues of state representation and function approximation. A formal assessment must address how cooperative behavior emerges from self-interested behavior, implications and limitations due to the sparse representation of action and the assumption of action independence, and the possible problem of individually nonstationary environments observed by each subagent. Crites (1994) reviews results in dynamic programming, automata theory, game theory, and other formal domains but finds few which seem to be directly relevant to questions of cooperative learning.

---

[1] Satinder Singh suggested exploring behavior in tasks which have Nash equilibria.

Chapter 8 Results of Motor Control and Integrated Model Simulations

In this chapter, we test HABLAR's motor control. We evaluate its ability to perform simple articulatory tasks without phonetic feedback. We also describe results of simulations which integrate audition with motor control, evaluating HABLAR's ability to learn a small inventory of vowels and consonant-vowel syllables. Planned simulations which test hierarchical aspects of HABLAR's motor control are also described.

8.1 Proximal task simulations

In this section, we test the effectiveness with which a single articulatory controller can learn proximal articulatory tasks in isolation without phonetic feedback. These are simple articulatory tasks measured directly from patterns of vocal tract and respiratory parameters, which attempt to shape articulatory patterns, not the resulting sound patterns. There are no phonetic targets, and the phonological controller does not generate reinforcement signals based on phonetic targets. This simplifies the articulatory controller's task and allows for easier diagnosis of failures. Distal tasks with phonetic targets and phonetic performance measures are the subject of Section 8.2.

8.1.1 Stop-consonant motion

Task 1 (reported originally in Markey 1994a) corresponds to the motion required to produce a simple stop consonant. Starting with a neutral, open oral cavity, the tongue must lift and make contact with a target region of the palate and then release contact, or the lips must close, touch, then release. The entire motion must occur in minimal time without encountering anatomically impossible configurations or making contact at a non-target location. Successful completion of the task within a time-out period garners a reward of 1.0 and ends the trial. An impossible physical configuration receives a penalty of –1.0 and ends the trial. Otherwise, a trial ends without reward or penalty after timing out. The set of gestures is somewhat different than listed in Table 7.3, involving 9 jaw angle, 5 lip height, 7 lip protrusion, 7 tongue tip angle, 6 tip length, 12 tongue body angle, 10 tongue body displacement, and 7 hyoid gesture targets. Three different target places-of-articulation are tested: labial, alveolar, and velar.

Examples of stop-consonant motion training sequences are pictured in Figure 8.1. In the figure's top three panels, contact of velum and release by tongue body is learned within 13,800 trials. During trial 200, every articulator moves — lips, tongue body and tip, hyoid, and jaw. By trial 13,800 motion is very economical, involving only tongue body and jaw; only the hyoid shows extraneous motion. This is an early simulation. More recent simulations pictured in the figure's bottom three panels and summarized in Table 8.1 learn these tasks much more quickly due to changes in how credit is assigned during the refractory period after a new gesture is initiated. In the bottom row of panels, contact of alveolar ridge and release by tongue tip is learned within 700 trials. In trial 200, nearly every articulator moves. In trial 500, the tongue attempts to lift toward the alveolar ridge, but falls short each time, finally succeeding after 358 msec. By now the controller has learned that the tongue body and tip are the salient articulators, but has not yet learned the entire motion. Trial 1900 accomplishes the task in 88 msec, without any extraneous motion.

We measure controller performance by the average time per trial necessary to complete the target trajectory and reach the final desired configuration. Worst case is to time out after 400 msec. Based on the time required for a hand-crafted solution to accomplish each task, best case is about 72 to 80 msec per trial.

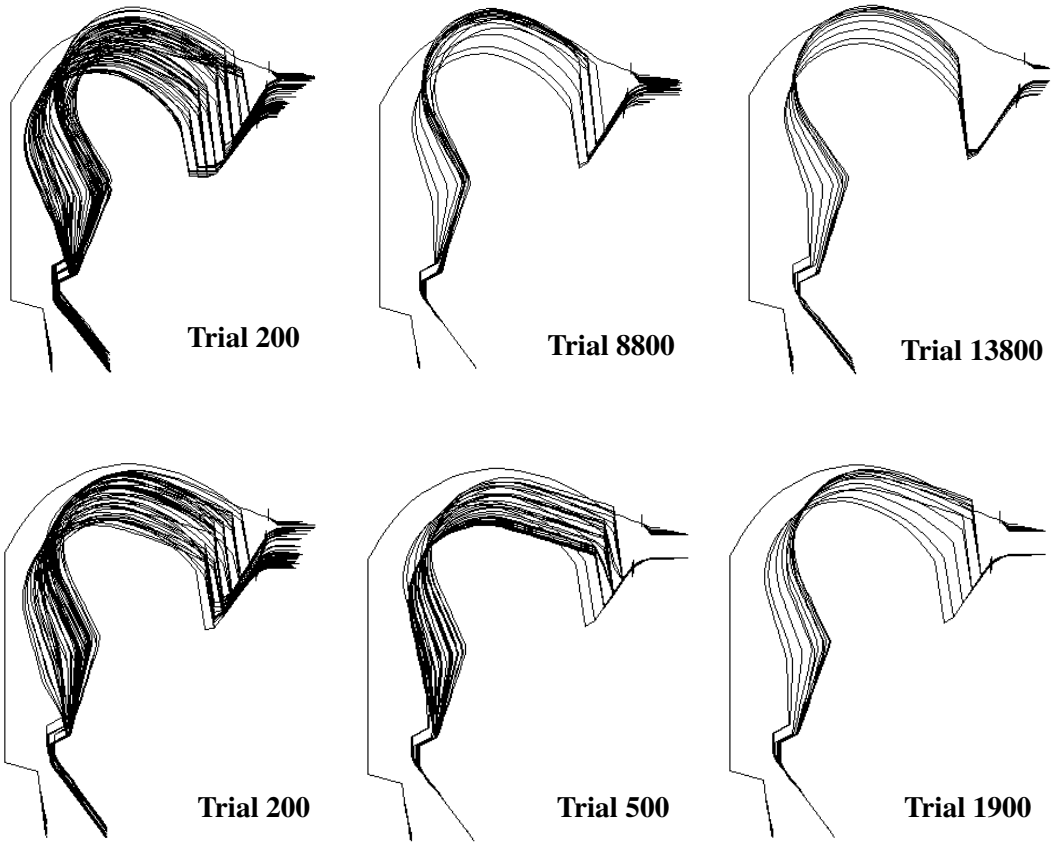


Figure 8.1 Articulatory controller learning consonant-like motion

In each panel, the vocal tract outline for each time step is overlaid, representing the entire range of motion for all vocal tract articulators for the duration of the trial from beginning until success or timeout. Lips are to the right; tongue body is to the left. See Figure 5.1 for the relative location of vocal tract articulators. Motion of the tongue tip reflects its own motion plus that of jaw and tongue body. However, it is possible to make out the range of tongue tip motions by looking for how its angle changes.

"Time" here is as simulated by the vocal tract model's dynamics. After as few as 700 trials, the articulatory controller has achieved near-optimal stop-consonant motions. Table 8.1 summarizes results.

Table 8.1 Stop-consonant contact and release

Place-of-Articulation	Time per trial after 10,000 trials	Trials to converge to 90% success	Model time to converge to 90% success
Labial	94.6 msec	860	4.2 minutes
Alveolar	116.4 msec	2810	9.9 minutes
Velar	92.5 msec	3610	12.8 minutes

8.1.2 Primitive consonant-vowel syllables. "Deaf" babble.

The second task calls for the generation of a primitive consonant-vowel syllable, targeting only the consonant as measured by place-of-articulation, but leaving the vowel unspecified. It starts with the motion required to make a stop-consonant but adds the requirements that the oral cavity remain open after the stop-consonant and that glottal vibration above a threshold intensity occur for at least 64 msec after articulator contact is released. Three different target places-of-articulation are tested: labial, alveolar, and velar. The inventory of gestures is similar to those listed in Table 7.3 except for fewer hyoid gestures, which are largely irrelevant to the task.

Strictly articulatory criteria are used to judge performance of this task, the tactile feedback of the tongue-palate or lips' contact, and tactile sensation of the glottal vibration. Acoustic feedback and phonetic characteristics of the resulting sound are ignored. Criteria for success are thus less exacting than if auditory perception were to test for some target syllable. In a sense, this simulates deaf babble which might occur if a deaf infant were to attempt a similar pattern of tactile feedback.

Average results for successful simulations and for all target places-of-articulation are summarized in Table 8.2. The average time to learn each task was only 42 minutes (in the model's simulated time scale). This seems reasonable even after adding additional respiratory time. Nine-thousand trials may seem quite high, but young children speak as many as 14,000 words per day (Wagner 1985). Training failures are discussed in Section 8.1.4.

Table 8.2 Articulatory controller learning of primitive syllables

Average number of trials to achieve 90% success	9,360 trials
Average time per trial	189 msec
Estimated "real world" time to learn each task	42 minutes

8.1.3 Reduplicated babble

In several instances, the sound generated by the primitive syllable simulations resembles reduplicated babble. This occurs when the articulatory controller approximates but does not achieve full articulator contact, then vocalizes and opens the oral cavity. Once open, the articulatory controller senses a state similar to the starting neutral configuration and attempts the stop-consonant once again. This can occur several times until full contact and subsequent release and vocalization are made. Once the target articulatory pattern is detected, the trial ends; however, if we allow the articulatory controller to continue once achieving its goal, reduplicated babble would continue until exhausting lung capacity.

8.1.4 Training failures

Simulations of tasks 1 and 2 sometimes fail. Once the articulatory controller initially learns the intended task, there is typically a period during which it is performed consistently. As training continues, performance sometimes deteriorates and the controller seems to unlearn the entire task. Upon closer inspection, these failures appear to be caused by generalization-based aliasing (see Section 7.4.3). Three factors contribute to the failures. First, tactile event memory and other temporal cues (e.g., gestural phase) are required to distinguish states before and after a consonantal release. Otherwise, rewards are not Markovian (see Section 7.4.4) and these tasks cannot be solved. Second, with its neural network implementation, it is not sufficient that the articulatory controller merely observe tactile event memory and other temporal cues. It must also abstract internal features which capture all functional distinctions among input states. Third, there is some evidence that internal features which encode recent tactile events are lost due to overtraining.

In these tasks, there are few constraints on the vocal tract's motion after consonantal release; the oral cavity must merely remain unobstructed. Thus, any sequence of gestures which opens the oral cavity and maintains continued vocalization after contact is acceptable. Often the gestures chosen by the articulatory controller after the consonantal closure take the tongue body *through a configuration nearly identical to the vocal tract's neutral open position*. After a period of additional training, the controller immediately seeks the post-consonant, fully open configuration without first attempting the consonant. The distinction between the neutral open configuration before the consonant and the neutral open configuration after the consonant seem to be lost. It is plausible that the controller overgeneralizes the final portion of its trajectory to the initial portion of its trajectory, ignoring state variables such as gestural phase and its memory of tactile contact which usually distinguish the two subtrajectories. This conjecture is strengthened by simulations which use a non-discrete representation of phase rather than the customary discrete, six-bit per articulator representation. The non-discrete representation failed eight times more often than the discrete representation.

These failures might be avoided by employing the experience replay method pioneered by Lin (1992). Unfortunately, articulatory state is too large and trials too long for this to be practical.

8.2 Integration of audition and motor control

The first step in evaluating HABLAR as an integrated sensorimotor system is to test the effectiveness of feedback from auditory perception in training articulatory controllers to pronounce simple target sounds. Several questions arise. (1) How robust is audition under less-than-ideal conditions? Simulations in Chapter 6 used well-behaved "parental" stimuli with well-defined syllables. HABLAR's untrained speech is not as well-behaved. Even if auditory perception cannot correctly classify all sounds, then can it at least provide feedback sufficient to guide early motor learning such that it becomes more speech-like? (2) How well does the articulatory controller scale up to more exacting articulatory problems than before? Simulations of stand-alone articulatory controllers above required the production of a primitive stop-consonant-vowel syllable, but tolerated any vowel. The constraints of speech are much narrower. (3) How effective is the phonetic distance metric and reinforcement function? The reinforcement function used to train articulatory controllers in stand-alone simulations was based on a proximal articulatory measure. In simulations reported in this section, reinforcement is calculated from a distal phonetic measure of success.

This initial set of simulations also served the purpose of calibrating model parameters (e.g., auditory, learning, reinforcement function, and various timing parameters and distance metrics) and correcting problems in the implementation.

8.2.1 Procedure

For these simulations, the model is configured with a single articulatory controller. Since there is no choice to make among articulatory controllers, the phonological controller's only purpose is to measure and transmit reinforcement signals to the sole articulatory controller.

Each trial begins with the presentation of a target stimulus. We assume that inhalation is a built-in vegetative function of the model, and pulmonary parameters are set to simulate lungs filled to tidal capacity. A trial ends after a maximum of 4 or 6 acoustic segments, 750 msec, or successful achievement of the goal, whichever comes first.

Articulatory controller rewards are calculated according to Equation (7.3) for each sequence of acoustic segments that ends with a static, prevoiced, or silent segment type. The articulatory controller receives a reward signal immediately after a stand-alone static segment or after a transient and static segment occur together, but not after a stand-alone transient segment. The purpose of this procedure is to correct for the disabled phonological controller and to reward entire syllables.

We present the model with several arbitrary stimuli to imitate, representing a range of targets of diverse articulatory complexity. These include isolated vowels /u/ and /&/, consonant-vowel syllables /bi/ and /bA/, and (other target sounds in progress). The /bi/ stimulus explicitly requires prevoicing before consonant release, but the /bA/ stimulus does not (and thus should be easier to achieve than the /bi/ stimulus).

We judge that the goal has been realized when the phonetic distance between actual and target utterance is less than 10% of the distance between the target and a null phonetic pattern. For these target utterances, this sufficed to identify those tokens whose segmental features matched the target exactly and whose spectral feature activations differed from the target by less than 0.5. At the conclusion of each trial, we also measure the cosine between actual and target phonetic representations for the entire utterance.

To evaluate results, we measure percent of successes, average phonetic proximity of actual and target phonetic representations, and cosine between actual and target utterance representations. We also determine the percent of trials during which the sequence of segment types matches the target, even if the spectral features do not match. Finally, we listen to samples of the sounds generated and transcribe them phonetically. This affords an unsystematic validation of the model's classification of its own speech.

8.2.2 Results

Of 24 preliminary simulations, each with one target utterance, 14 result in stable or quasi-stable babble or speech-like sounds, of which eight were relatively close to the target sounds. Table 8.3 presents a summary of select simulations. These are *initial results*, run with a variety of *previously untested* training parameter settings.

Most simulations converge quickly on the broad properties of the target sounds, mastering voice control, achieving stable or quasi-stable speech-like sounds, and exhibiting phonetic features present in the target sound, even if the target sound was never accurately replicated. Based on our transcription of its sounds, the model's phonetic classification of its own speech appears to be relatively accurate. Most importantly, these initial results suggest that auditory feedback is an effective reward signal.

Most problems appear to involve motor control. Simulations often converge on local optima or oscillate among suboptimal solutions. Some such local optima are the result of anomalies in the model's calculation of phonetic distance, but this should be easily corrected. We also suspect that generalization-based aliasing similar to that observed in Section 8.1.4 affects some simulations, but we have not yet confirmed this with a complete analysis.

Table 8.3 Select integrated sensorimotor simulations

Target Sound	Observed Sound	When observed	Cosine	Comments
u	[A]	300 trials	0.724	Very stable but suboptimal pattern.
u	[u] or [U]	2300 trials	0.979	Sound is shorter than a true /u/, sounding closer to /U/, because trial ends as soon as the static segment is detected.
u	[hAjAjAjA] and similar.	8000-11000 trials	0.463	Quasi-stable babble-like pattern.
u	[hU, huE, huB, huF]	1000-4000 trials	0.501	Considerable variation in productions, but consistent presence of [u] sounds, preceded by aspiration [h] and usually followed by fricative sounds.
&	between [&] and [E]	6000 trials	0.765	
&	[hAe, hoA, hA, huI, hAE, huE]	2000-4000 trials	maximum of 0.500	Never converges. Final segment in many sounds is phonetically close to & (e.g., E and e).
bi	[dubbub, budbub], etc.	11000 trials	0.328 to 0.428	Never converges, but about 15-20% of trials are true stop consonants, matching the target's prevoicing-transition-static segment type pattern.
bA	[bi]	4700 trials	0.842	
bA	[hAwA, hA^wA, hAUA]	6000-9000 trials	0.735	wA pattern is consistent with /bA/ without prevoicing.
bA	[bA] or [wA] without plosion	5100 trials	0.808	Sounds like bA or wA without the prevoicing and plosion
bA	[bI]	5400 trials	0.680	Not stable.

Thus, performance in these initial simulations is imperfect. But children's speech is imperfect, too. The long term aim of this research (see Chapter 10) is to determine whether the errors made by HABLAR correspond to those made by children.

8.3 An engineered hierarchy

The purpose of the next set of simulations is to determine HABLAR's ability to compose simple sequences of sounds using an engineered hierarchy. This will evaluate the compositional characteristics of the phonetic representation, the effectiveness of phonetic distance and reward signals, and adequacy of the phonological controller's learning algorithm.

We train the model to imitate targets composed of sounds learned by articulatory controllers in the simulations reported in Section 8.2. The model is configured with several articulatory controllers, each initialized with weights saved from the previous simulations. Articulatory controller weights are frozen (or the learning rate and temperature are set very low), and only phonological controller weights are allowed to vary.

8.4 Phased complexity simulations

As a final demonstration of the model, we present it with the task of learning to "imitate" nine CV syllables built from three consonants (b, d, g) and three vowels (i, a, U). Veridical phonetic representations of each target utterance are presented as "parental" stimuli. The goal is to generate sounds whose phonetic representations are as close to target utterances as possible. Parental stimuli are selected from a set of tokens already built to test auditory perception.

The model is fully configured, a test of the fully integrated sensorimotor system, including the phonological controller's ability to guide the acquisition of sounds which approximates the distribution of sounds in the environment under different bootstrapping strategies.

We vary the order of stimulus presentation to simulate different simple bootstrapping strategies: (1) random (no bootstrapping), and (2) extended periods of training with each syllable type. We also vary the number of articulatory controllers, and relative training and annealing rates of articulatory and phonological controllers.

Each trial begins with the presentation of a target stimulus. We assume that inhalation is a built-in vegetative function of the model, and pulmonary parameters are set to simulate lungs filled to tidal capacity. Each trial ends (1) when the distance between target and feedback phonetic representations falls within some tolerance, (2) when the distance rises beyond some threshold, or (3) after a time-out period.

Based on simulations to date, we expect that the integrated model will initially demonstrate poor articulatory control and considerable variation with a large number of non-speech sounds. It will show evidence of increasingly competent control but randomly distributed syllable types resembling random but canonical babbling (Oller & Lynch 1992). The distribution of syllables will initially reflect the natural constraints of vocal tract anatomy, showing a wide variety of vowels and stop consonants. But after a time, the distribution of syllables will start to approximate the distribution of syllables in "parental" stimuli, resembling the process of accommodation which occurs in infant babble (Vihman et al. 1986).

Chapter 9 Explaining Developmental Psycholinguistic Data

Results of simulations with HABLAR conducted to date demonstrate the faithfulness of its auditory perception, the feasibility of articulatory control in many dimensions, and the effectiveness of integrating auditory perception and articulatory control modules. It will take time to systematically and exhaustively test the model's fidelity with respect to developmental phenomena. A methodology for this task is presented in Chapter 10 (see Section 10.1). In the meantime, it is reasonable to ask if the model can plausibly simulate or explain developmental phenomena. In this chapter, we examine HABLAR's properties and explore the model's qualitative explanations of developmental phenomena. We also identify its weaknesses and how they might be corrected.

9.1 HABLAR's properties which contribute to observed phenomena

In the introductory chapters, we drew from observations of phonological phenomena a number of key abstractions which constrain HABLAR's design. These include the assumption that some veridical learned representation of speech underlies children's pronunciation. This representation classifies local, static, contextual, and dynamical acoustic features in a way which captures the phonetic similarities among sounds and their relative order. Associative, stochastic mechanisms are assumed to link this internal representation of target sounds and their articulation. Anatomical constraints and motor skill are assumed to play a central role in the evolution of the model's phonological competence. Finally, the dual compositional and articulatory character of speech motivate the model's form of hierarchical motor control.

Having built HABLAR based on these constraints, we now seek to identify the model's emergent properties, how its components behave and interact as implemented, and how these characteristics explain developmental phenomena. Lacking exhaustive simulations, we hypothesize how the integrated model will behave based on observations of partial simulations, architectural relationships among model components, and known properties of connectionist learning systems and other mechanisms used in HABLAR's construction. We expect that several mechanisms contribute to each behavioral pattern. For example, anatomical constraints, acoustic properties, proprioceptive cues, and properties of the model's learning algorithms all contribute in various ways to substitution errors. Some mechanisms seem to contribute to multiple phenomena. Rather than explaining them repeatedly, we summarize them in this section. The following sections then explain key data using these and other model properties.

9.1.1 Effects of articulatory mechanisms

Some sounds are more difficult to make than others. In HABLAR this is expressed as the relative likelihood of generating each sound, which is a function of several mechanisms: anatomical, acoustic, and aerodynamic constraints, the sound's articulatory complexity, and the availability of proprioceptive cues. For example, a fricative occurs only within a small range of vocal tract configurations and requires more precise control of articulatory motion than a related stop consonant. The probability of generating each sound is also a function of articulatory controller weights and other parameters (e.g., Boltzmann distribution temperature). At the beginning of training, the weights of each articulatory controller are randomly initialized and temperature is high, leading to a distribution of sounds which reflect on average the priors of anatomy, acoustics, and aerodynamics. These factors continue to bias the distribution of speech sounds and contribute to sound substitutions even after weights have been shaped by training.

Proprioceptive feedback includes information about past tactile and respiratory events. However, memory of such past states is limited. Consequently, the articulatory controller's ability to distinguish between states distant in time but otherwise alike in proprioceptive feedback — hence its ability to distinguish between different sequences of articulatory events — is also limited. Given the articulatory controller's closed-loop control, one likely result is repetitive articulatory motion, including reduplicative babble, a result observed in some preliminary simulations (see Section 8.1.3).

9.1.2 Effects of stochastic gradient descent learning algorithms

HABLAR assumes only scalar rewards and costs. Without a signed error vector, the model relies on stochastic reinforcement learning methods to evaluate the relative utility of different actions and to correct its utility estimates based on reinforcement signals it observes. After an action is taken, each controller incrementally changes weights to reduce its error in predicting future rewards. Errors are reduced in the direction of greatest error reduction, i.e., in the direction of the error gradient, or equivalently in the direction of incrementally greater rewards. In order to explore the consequences of all actions, suboptimal actions are sometimes chosen, although actions which are expected to maximize future rewards are more likely to be selected (see Section 7.6).

Thus, each controller engages in a form of stochastic gradient descent learning. At each control level there is a gradual, stochastic approximation of target pronunciations. The gradient descent character of the process ensures that the biggest errors get eliminated first. The stochastic character of the process ensures individual behavioral differences. It also results in trial-to-trial behavioral differences which are observed as phonetic variation when they occur at the articulatory control level and as compositional alternation when they occur at the phonological control level (see Section 2.7).

Pronunciation errors and substitutions are in part a function of articulatory difficulties in producing various sounds, initially inaccurate estimates of future rewards, the gradual correction of estimated utilities, and random suboptimal actions.

9.1.3 Effects of the phonological controller's associative architecture

The phonological controller's associative mapping of phonetic targets to articulatory sequences is subject to generalization across phonetic similarities. This derives from two sources: (1) basic generalization properties of its neural network implementation, and (2) phonetic similarities captured in target sound representations formed by auditory perception. These properties contribute to the phonological compositionality of the model's utterances (Section 2.3.1), but it also ensures overgeneralization errors (Section 2.5). Associative models are also subject to overtraining and memorization, contributing to the inertia observed when particularly old pronunciation patterns resist change (Section 2.5).

9.1.4 Effects of external and internal phonetic features

All phonetic features in a syllabic target are present in the target phonetic representation, and all are present as inputs to the phonological controller. Hence, all affect the output. This makes possible phenomena in which the target word's component sounds get rearranged or transposed (e.g., metathesis) and in which non-contiguous target sounds affect each other (e.g., consonant harmony: see Section 2.4). Contextual features in the representation also contribute to context sensitive errors (see Section 2.2).

Internal subsymbolic cues learned gradually from the associative mapping between phonetic features and articulatory events govern utterances more directly than features in the target word's phonetic representation. In isolation, a target sound's phonetic features have no articulatory significance, only

acoustic and statistical distinctiveness. The articulatory significance of phonetic patterns must be learned. This is the role of the phonological controller as it learns which articulatory controller is most likely to generate each target sound. As it does so, it develops weights which encode phonetic-articulatory relationships and internal distributed representations (expressed as activation patterns of its hidden units) which signify the articulatory determinants of the phonetic target.

The subsymbolic, distributed nature of these internal cues and the incremental manner with which they are learned contribute to irregularities in developing speech (Section 2.6). It also contributes to the systematicity of pronunciation patterns and errors, especially the systematicity of context sensitive patterns such as consonant harmony (see Section 9.2.6, below).

9.1.5 Effects of the phonological controller's model of speech sounds

HABLAR differs from simple articulatory skill models (Section 3.4) by its introduction of a phonological control level between auditory perception and articulatory control. Superficially, it seems to possess some attributes of abstract phonological models (Section 3.3), composing phonological patterns from elemental phonetic units. From a somewhat different perspective, it seems to function as a sequential gating mechanism, selecting the sequence of articulatory controllers best suited to generate a target sound. But its role is more abstract, involving more than motor control. Together with articulatory controllers, the phonological controller builds a statistical model of the likelihood with which each articulatory controller generates some sound. This statistical model is absolutely fundamental to HABLAR's behavior. It serves the same function as an inverse model. Substitutions occur when the statistical model is wrong. Changes contribute to the evolving phonetic characteristics of babble and speech. It may also explain the direct perception of articulatory gestures (Fowler & Rosenblum 1991) and the apparent gestural character of speech perception (Liberman & Mattingly 1985; see Section 9.4.1). It resembles a statistical mixture model.

A *statistical mixture model* is a set of statistical distributions whose parameters may be adjusted to simulate a set of environmental observations, especially observations whose values are clustered into groups. An example is the Gaussian mixture model (Nowlan 1991a,b), which is used by HABLAR's auditory perception to categorize static and dynamic spectra (see Section 6.2.5 and Figure 6.2). A Gaussian mixture model's goal is to simulate a set of observations with not one but many Gaussian distributions. Given an observation it determines which of several Gaussians is most likely to have generated it, calculating the conditional probabilities for each. The mixture model is updated by incrementally changing the means and variances of each Gaussian in proportion to the likelihood that it generated the observation.

In HABLAR, articulatory controllers correspond to the statistical mixture's multiple distributions. Each target utterance corresponds to an "observation." The phonological controller estimates the likelihood with which each articulatory controller (distribution) generated the target utterance (the observation). This is pictured in Figure 9.1. Each possible sound is produced with some probability by each articulatory controller (as shown schematically in the top portion of the figure). Given a target utterance, the phonological controller asks which articulatory controller is likely to have generated it — not by computing conditional probabilities directly, but indirectly by computing the Boltzmann distribution over its Q-values per Equations (7.7) and (7.4). The Boltzmann distribution is the mechanism by which a Q-agent chooses an action and experiments with optimal and suboptimal choices. Having been chosen according to this distribution, an articulatory controller becomes active and its behavior is observed. Unlike a Gaussian mixture model, there is no simple "mean" and "variance" to adjust. The articulatory controllers' parameters are updated indirectly by reinforcement learning. The chosen articulatory controller's "mean" sound stochastically and incrementally moves closer to the target sound (as shown in the figure). The phonological controller adjusts its model of articulatory controller behavior, too.

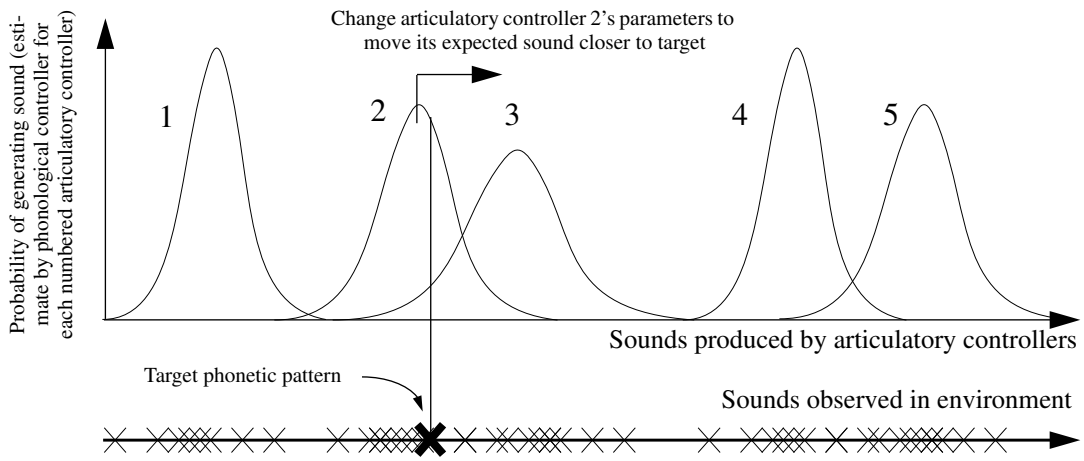


Figure 9.1 A statistical mixture model of the linguistic environment

9.1.6 Other effects of computational mechanisms

A number of other computational attributes affect HABLAR's behavior. The nonstationary environment observed by the phonological controller and the computational complexity of hierarchical motor control contributes to the difficulty of learning phonological control. On the other hand, excess articulatory degrees-of-freedom together with the wide acoustic latitude afforded sounds in each phonetic category by HABLAR's categorical perception make possible a wide range of articulatory solutions for each target sound.

9.2 Qualitative Explanations and Predictions

Having characterized key mechanisms which contribute to pronunciation errors, we now explore several specific phenomena in more detail, showing how these mechanisms and their properties interact to generate observed behavior.

9.2.1 The phonetic characteristics of babble

Children's earliest babble shows a universal preference for sounds which do not necessarily match the statistics of adult speech (e.g., Kent & Bauer 1985, Locke 1983, Oller et al. 1976). 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 (de Boysson-Bardies et al. 1992, de Boysson-Bardies & Vihman 1991, Vihman et al. 1986, Whalen et al. 1991).

HABLAR's explanation draws on three principal factors: the relative difficulty and likelihood of producing various sounds, evolving articulatory controller skills, and the phonological controller's evolving map between phonetic targets and articulatory events.

We assume that speech targets drawn from imitation and lexical memory reflect the distribution of sounds in the environment. Motor skill limits the sounds actually generated. With minimal experience, the distribution of sounds generated by articulatory controllers will show high variance and will reflect relative articulatory difficulty (and contributing factors such as anatomical and acoustic constraints, articulatory degrees of freedom, and proprioceptive cues) rather than linguistic constraints. Likewise, the phonological

controller's estimate of articulatory controllers matched with target sounds will reflect this. However, with greater experience and guided by rewards which favor phonetic targets perceived in the linguistic environment, articulatory controller behavior narrows and comes to reflect linguistic requirements. Also, the phonological controller will more accurately model the mapping between target sounds and articulatory controller capabilities. This merely restates the process which occurs as HABLAR's motor control builds a statistical mixture model of linguistically relevant targets (see Section 9.1.5). These factors ensure that vowels, nasals, glottal stops, and primitive stop-vowel syllables will occur most frequently in early HABLAR utterances, precisely the vocalizations most frequent in early babble.

We have not yet completely assessed the characteristics of HABLAR's babble, but a crude comparison suggests that its initial behavior parallels early babbling by children, governed more by anatomical constraints than by the linguistic environment. Table 9.1 tabulates HABLAR's vocal tract constrictions by place-of-articulation during the first 100 trials of 40 simulations. It also tabulates the distribution of stop-consonant places-of-articulation observed in infant babble before any active vocabulary has been acquired (Vihman et al. 1986). The two distributions are qualitatively alike. Differences from human data are unavoidable, since HABLAR's anatomical and other constraints do not precisely match those of an infant. The table reports percent at each place-of-articulation, with standard deviations in parentheses.

Table 9.1 Infant stop consonants vs. HABLAR vocal tract constrictions

Subject	Labial	Dental	Velar
Infants during babble (when active vocabulary is 0 words)	44% (26)	43% (26)	13% (9)
HABLAR (during 1st 100 trials, average of 40 simulations)	34% (11)	35% (12)	23% (12)

9.2.2 Sound substitution

Perhaps the most characteristic pronunciation error is children's substitution of a target sound with a phonetically similar but easier to pronounce sound: fricatives with stops, liquids with glides, voiceless word-initial consonants with voiced consonants (see Section 2.1).

In HABLAR, producing some sounds is more probable than others due to the anatomic, acoustic, and aerodynamic constraints and their interaction, relative articulatory complexity, and proprioceptive cues available to guide articulatory control (see Section 9.1.1). For example, compared with fricatives, stop consonants require less precise control, rely on fewer, clearer proprioceptive cues, and can take greater advantage of the vocal tract's excess degrees-of-freedom. Until mastered, the likelihood of accurately rendering a fricative is lower than that of a phonetically similar stop consonant.

When choosing between two pronunciations (e.g., [T] vs. [t]), the phonological controller's choice of articulatory controller depends on reward probabilities. It will weigh the much higher probability of receiving a somewhat lower reward (e.g., substituting a reliably produced [t] for the target sound [T]) against the much lower probability of receiving a higher reward (e.g., choosing an articulatory specialist which may produce [T], but unreliably). This is illustrated in Figure 9.4.

Hence, HABLAR will usually substitute a less accurate sound (e.g., pronouncing *thank* [T&Nk] as "tank" [t&Nk]) or omit a sound segment if the estimated *net* reward for the resulting utterance is higher. Recall that the phonological controller is rewarded based on the accuracy of the *whole word*, as determined by the phonetic distance between target and actual words. The phonetic distance between *thank* and *tank* is small, as their representations differ only in the type of onset segment detected in each word (e.g., static fricative vs. burst segments). Even their spectral onsets (e.g., [T&] vs. [t&]) are similar.

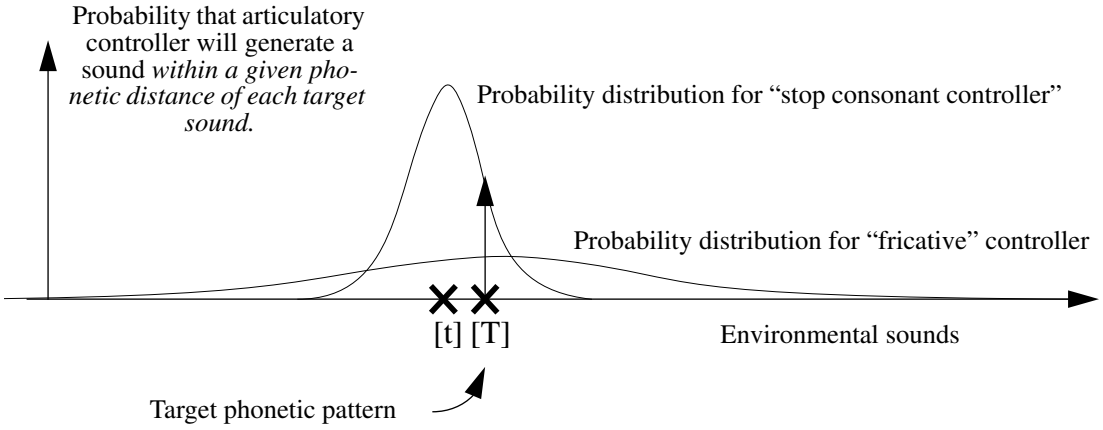


Figure 9.2 Sound substitution

Substitution occurs when the likelihood of reward due to an inaccurate sound exceeds likelihood of reward due to an accurate sound. Probability distributions portray the likelihood of generating a sound within some phonetic tolerance of the target.

9.2.3 Sound competition

As error patterns change, they give way to newer errors or accurate, adult-like pronunciations. Often they do so only gradually. During this period old and new patterns compete. Contributing are the same mechanisms which underlie sound substitution, plus HABLAR's gradual, stochastic approximation of target pronunciations (see Section 9.1.2).

The model always computes the relative likelihoods that different actions produce the target sound. Often the relative likelihood of one sequence of actions will greatly surpass all others. If not, however, the actions and their phonetic consequences will compete. More technically, the likelihood of each action's generating a sound is explicitly determined by computing the Boltzmann distribution over the Q-values for all possible actions. If the Q-values of two actions are equal, they and their subsequent trajectories are equiprobable.

Actions may compete at the articulatory or phonological control levels. The phonological controller is responsible for alternation of larger chunks of sound whose specific articulatory details have been mastered (e.g., Daniel's harmony patterns, see Table 2.4 on page 11). Phonetic variation of yet unmastered sound segments is the responsibility of the articulatory controller, although phonetic variation and compositional alternation may often overlap.

Linguistically, is this form of competition among sounds free variation or alternation? In adult phonology, alternation denotes the relationship that exists among alternative forms or variants (Crystal 1985), conditioned on some phonetic, morphological, or grammatical context. According to HABLAR's account, the variation is statistically conditioned on subsymbolic features in phonetic context; hence, it is alternation.

9.2.4 Voicing contrast

Macken & Barton (1979) analyze the acquisition of the voicing contrast in considerable detail for 4 children between 18 to 25 months, based not only on longitudinal phonetic transcriptions but acoustic measurements of voice onset time (VOT) as well. They identify three overall stages: (1) no contrast, (2)

apparent acoustic contrast but not consistently perceptible to adults, and (3) contrast which resembles adult speech. Stage two is characterized by a gradual statistical approximation of adult VOT. Thus, children seem to recognize the contrast even if they have not yet mastered it in speech production. This observation supports HABLAR's assumption that a veridical representation of target sounds underlies children's speech. The observed gradual statistical approximation of adult VOT is what would be expected from HABLAR's stochastic gradient descent.

The shift between stages two and three is abrupt. Early in stage three, aspiration (which is necessary to achieve target VOT durations) is employed for voiceless stops for the first time, but mean VOT greatly exceeds mean adult VOT, and VOT distributions are relatively uniform. Only later in stage three do VOT distributions more closely approximate the adult target. When aspiration is first introduced, it is as if there were no information about the magnitude of error correction needed to achieve the target VOT. The pattern seems plausible for a learning paradigm in which articulatory innovations are introduced by random experimentation and in which there is no signed error vector. Certainly, however, this conjecture must be confirmed by simulation.

9.2.5 Context sensitive generalization

Children's pronunciation errors are special because of their context sensitivity (see Section 2.2) and because of the approximately lawful relationship between target and actual utterances (Section 2.3). Many such errors emerge when newly learned sounds (often accurate at first) are overgeneralized to inappropriate lexical contexts. For example (see Section 2.5), Daniel started to use word-initial nasals innocently enough in his nearly correct pronunciation of *moon* [mun, mum], but then inappropriately applied initial nasals to *broom, mug*, *going*, *spoon*, and *prune* (Menn 1971, unpublished data).

In HABLAR, the phonological controller generalizes by producing similar articulatory patterns in response to similar phonetic representations. Phonetic representations tend to be similar for acoustically similar words. Moreover, the phonological controller's input spans the whole target syllable, and phonetic features capture the coarticulation of contiguous sounds. Thus, HABLAR's generalization patterns are expected to be context sensitive in ways which differ from adult phonotactic rules and in ways which resemble child-like context sensitive errors. For example, in adult targets, the vowel preceding a nasal is also nasalized (e.g., *tune* [tũn]). In HABLAR, the nasalized vowel will affect the phonetic representation of the onset demisyllable (e.g., [tũ]) in consonant-vowel-nasal targets. This may cause the target's initial consonant to be pronounced as a nasal instead of an oral stop, explaining some forms of consonant harmony as in the example above (see also Levelt 1994).

9.2.6 Consonant harmony

Consonant harmony is one of the more dramatic types of context sensitive errors. It often cannot be explained by the model's tendency to capture the contextual phonetic properties of contiguous sounds, nor by any other single factor. Instead, a plausible story emerges from contributions of two factors: the behavior of HABLAR's gradient descent learning mechanisms and the uneven rate with which the phonological controller learns the articulatory correlates of phonetic features in its input.

This story is best told by an example: the substitution of *guck* [gAk^{h}] for *duck* [dAk^{h}]. The relevant mechanisms and their interaction are illustrated in Figure 9.3. The top portion of the figure portrays the phonetic features activated by the target utterance [dAk^{h}]. Highlighted are the most salient groups of phonetic features, those which act as cues for syllable shape and place-of-articulation. Phonological controller mechanisms are explained in the middle portion of the figure. The bottom portion displays the phonetic features of the actual utterance. Details of articulatory controllers are omitted.

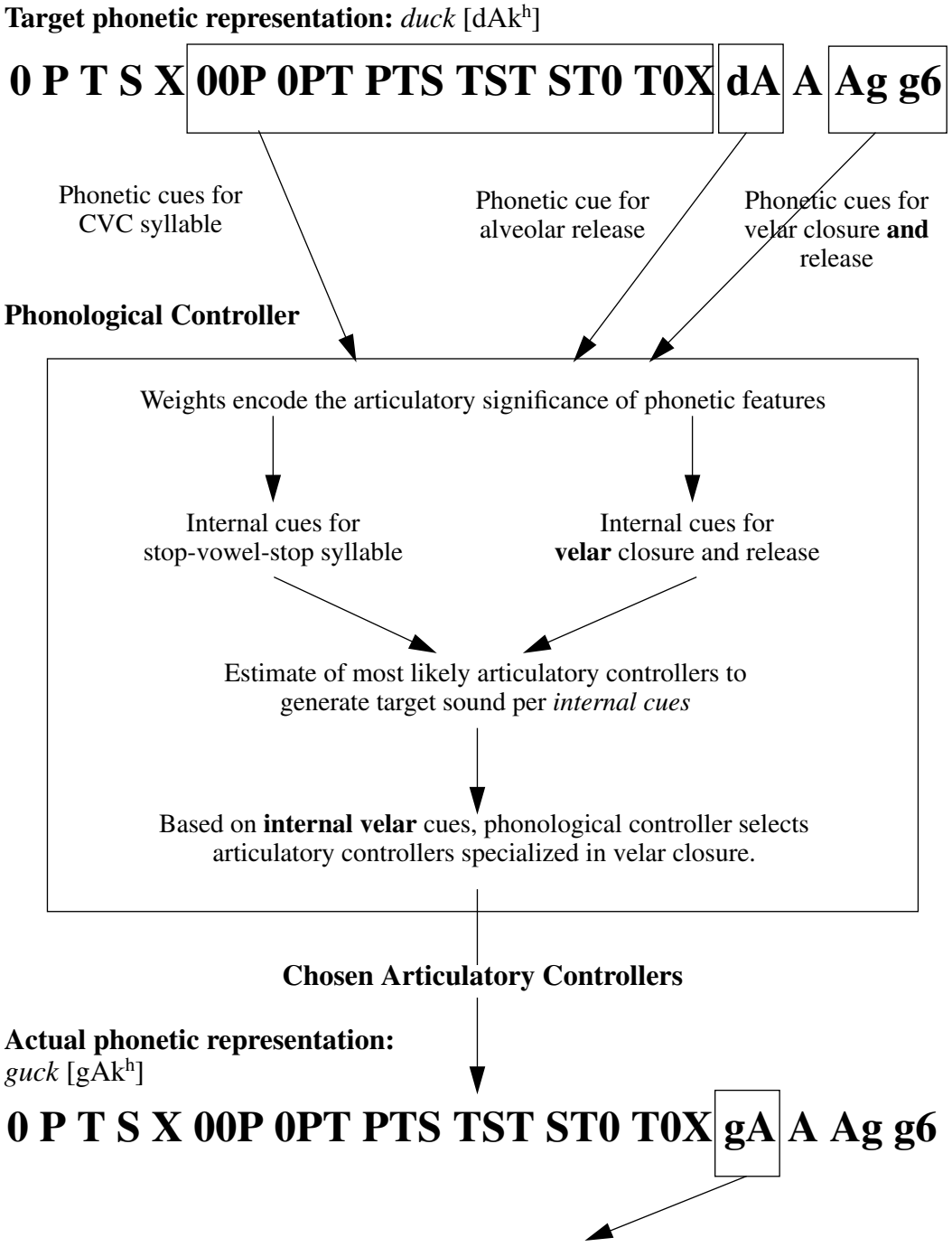


Figure 9.3 Consonant harmony — how *duck* becomes *guck*

HABLAR's gradient descent based learning algorithms tend to eliminate the biggest sources of error first (see Section 9.1.2). The phonetic representation of an utterance affected by consonant harmony often differs by a *single feature* from the accurate target representation and is much closer to the target than many other possible pronunciations. For example, the representation of *guck* differs by only one feature from the target *duck* (e.g., [gA] vs. [dA]). Table 9.2 summarizes the squared Euclidean distances between the target and various CV and CVC forms (assigning a feature activation of 1.0 for each segmental and spectral feature). *Guck* is closer to *duck* than virtually any other form.

Table 9.2 Phonetic distance between target and various pronunciations

Alternative forms for *duck*	**Square distance from target**
dAk^h (target)	0
gAk^h, bAk^h	2
dik^h, dAt^h	4
dA, dAx, gAg	8
dAd	9
bA, dAn, dAN, gA	10
nAk^h	11

Unfortunately, other sounds are just as close to the target as the harmonic form, or nearly so (e.g., [bAk^h, dAt^h, dik^h]). Moreover, phonetic distance does not distinguish consonant harmony from simple forms of substitution. *Guck* is the same phonetic distance from *duck* whether [g] replaces the adult /d/ in all contexts or only when followed by a word-final velar. Recall that harmony's presence is betrayed by its *systematic context sensitive distribution*, not by individual words. Daniel's *guck* is considered consonant harmony and not simple substitution because open /dV/ syllables are pronounced accurately (e.g., *tea* [di]) but [g] replaces initial consonants in all velar-final forms regardless of the initial consonant (Menn 1971).

Context sensitivity is the distinguishing property of consonant harmony. Non-contiguous sounds in the target affect each other in the child's output. The final velar in *duck* or *book* colors the place-of-articulation of initial alveolar or labial consonants. We must look for mechanisms by which this occurs. We argued in Section 9.1.4 that the model's whole-syllable phonetic representation enables the rearrangement or assimilation of sounds and other context sensitive effects. Certainly, that is the case here. The *entire* target phonetic representation and *all its phonetic features* are present in the input to the phonological controller (as shown at the top of Figure 9.3). It contains cues for the onset's alveolar release [dA], the coda's velar closure [Ag] and velar release [g6]. Unfortunately, this is necessary but not sufficient to cause the systematic occurrence of consonant harmony observed in some children (Menn 1971, Smith 1973, Vihman 1978; see Section 2.4). A mechanism is needed by which harmony is generalized in the right contexts.

A complete account requires an understanding of the mapping performed by the phonological controller. Its input is the activation pattern of all phonetic features detected in the target syllable, including those associated with the initial and final consonants. However, phonetic features only encode the distinctiveness of sounds, not their articulatory correlates. Articulatory correlates are acquired only when the phonological controller learns which articulatory controller is best suited to generate the target sound. Articulatory correlates are therefore encoded in the phonological controller's weights. This mapping is learned gradually. It is quite possible at some intermediate stage while learning the complicated mappings of consonant-vowel-consonant syllables, that the phonological controller encodes the articulatory correlates of some but not all phonetic features in the input. It might ignore the place-of-articulation of one consonant but not another, yet accurately encode the basic shape of the syllable.

Indeed, there is no guarantee that all articulatory correlates of the input's phonetic features will be learned at the same time. Properties of the phonological controller's gradient descent learning algorithm imply just the opposite — that the biggest sources of error will be eliminated first. For example, phonetic features in *duck* which signify the overall syllable shape far outweigh features implying place-of-articulation. Triads TST, ST0, and T0X and the simultaneous presence of dynamic spectra [dA], [Ag], and [g6] distinguish the target's CVC shape from the shape of a simple CV syllable. The phonological controller's learning algorithm is likely to solve the problem of syllabic shape first, even at the expense of some phonetic details like place-of-articulation.

In our example, the phonological controller might learn the articulatory correlates of [Ag] and [g6] phonetic features: velar closure and release. It might also learn that two vocal tract closures and one stop consonant release are implied when PST, TST, ST0, and T0X prosodic triads are in the target's phonetic representation. However, it might not immediately or completely learn the alveolar articulatory correlates of the other dynamic spectral cue, [dA]. Thus, the activation of velar and stop-vowel-stop phonetic cues would far outweigh the single phonetic cue for the syllable's alveolar onset, affecting the choice of articulatory controllers. In contrast, if no velar phonetic cue were present in the target word (e.g., Daniel's *tea*), then alveolar cues would not be outweighed by velar cues, and the word would be accurately rendered.

Finally, the model's hidden unit weights encode the articulatory significance of phonetic features in the target utterance. This enables generalization of consonant harmony to other stop-vowel-stop, velar-final targets. Thus, *guck* is pronounced instead of the target *duck*, *gook* [gUk] for *book*, *gink* [gInk] for *drink*, and so on (Menn 1971).

The model makes some tentative predictions about consonant harmony: its stability, the possible effect of unrelated phonetic cues, and the relative frequency of progressive and regressive harmony.

Depending on the model's tolerance for error, a harmonic form will obtain a smaller reward, but a reward nonetheless. It will remain stable unless there is a more accurate competing form. There is no systematic developmental evidence in support of this conjecture, but some anecdotal evidence exists. For example, Daniel's velar harmony lasted for nearly a year before competing forms appeared (Menn 1971, unpublished data). Once introduced, however, the accurate competing forms quickly became dominant and velar harmony disappeared (Menn unpublished data, see Table 2.4 on page 11).

According to this account, output forms are a function of all phonetic input cues. Thus, consonant harmony may be conditioned on the presence or absence of other seemingly irrelevant phonetic features in the target word, including vocalic context. Although such effects will be idiosyncratic, we find several well-analyzed cases to support this contention. Amahl harmonizes nasals and continuants but not stops (Smith 1973). Daniel does not always harmonize consonants for place-of-articulation if one consonant or the other is nasal (Menn 1971, unpublished data). For a number of children learning Dutch (Levelt 1994), consonant harmony appears to be conditioned on the vowel between the consonants.

The role of gradient descent learning in this account suggests that harmony is more likely between phonetically similar segments (i.e., segments which differ by only one feature). Otherwise, the distance between target and harmonized form would be greater. This is indirectly supported by Vihman's (1978) finding that full harmony (assimilation of all features) is more frequent than partial harmony. Also, Daniel harmonizes place-of-articulation among consonants which differ only in place or perhaps voicing (Menn 1971). However, some of these observations may be explained by other influences. For example, fricatives may be entirely deleted and thus not come under the influence of consonant harmony.

HABLAR also makes a weak prediction with regard to the relative frequency of progressive and regressive harmony for $C_1VC_2^h$ and $C_1VC_2V...$ target words. Onset consonants typically only have a single phonetic cue, the onset transition (C_1V). Medial and coda consonants have multiple cues: the transition during closure (VC_2) and the transition during release for coda consonants (C_2^h) or the transition into the

following vowel for medial consonants (C_2V). Holding other factors constant, multiple spectral cues may give coda and medial consonants an edge in developing articulatory associations. This should favor the occurrence of regressive consonant harmony. Of course, other factors, particularly the relative salience of onset transitions, coda transitions, and consonant releases, the strength of medial syllables, morphophonemic influences, and adult phonotactics all affect the strength of these cues. Vihman's (1978) analysis supports the prediction. However, the analysis is a compilation of several longitudinal studies without controls for linguistic environment and without methodological consistency among the studies compiled.

Surely, this section on consonant harmony is speculative, and its conjectures have not yet been demonstrated with quantitative simulations. A evaluative methodology is summarized in Chapter 10. More generally, HABLAR offers an experimental paradigm with which to make testable predictions about consonant harmony and other phenomena for specific linguistic environments and specific language-learner strategies. It should be possible, for example, to test the effect of articulatory patterns favored during babble. Menn (1983) suggests that differences in tolerance for phonetic inaccuracy may influence the types of errors children encounter. This too should be experimentally testable.

9.2.7 Pronunciation difficulties and learning strategies

Children use an amazing array of strategies in learning how to pronounce words, differing in their phonetic preferences, tolerance for error, and developmental paths (Menn 1983, Vihman 1993). What is the ideal strategy? Perhaps an analytic, gradual approach which avoids targets too distant from already-mastered sounds and which deliberately paces the increase in target complexity.

To our knowledge, no definitive studies have been conducted of the efficacy of different acquisition strategies, but in a reanalysis of several children's data, Peters (1977) observes that children with an "analytic" or "referential" strategy, who limit their early utterances to single words (or perhaps a subset thereof), are easier to understand. Children with "global" or "gestalt" strategies, who attempt multisyllabic words or whole phrases and sentences prematurely, are difficult to understand and show much more phonetic variation. She terms this latter style of speech "mushmouth."

Observation of children prematurely attempting to master individual complex targets reveals similar difficulties. One of Jacob's first words is *thankyou*. He used it to request objects or assistance before mastering its component sounds. Jacob experiments with an untold number of variations for six months before finally giving up, never once accurately rendering the complex utterance (Menn data 1976). Some features match the target. Attempts usually have two-syllables with dental and velar consonants. First syllable vowels are typically front, but second syllable vowels are back or central and range between middle and high. However, the sequence, voicing, and manner of consonants are never mastered. For example, Jacob experimented with all possible dental and velar consonant sequences [deigA, dido, d&tA, geigu, gigo, gi:do]. Figure 9.4 summarizes *thankyou*'s place-of-articulation variation between 12 and 18 months.

Initial difficulties in mastering elementary sounds may in part explain the long period of seemingly aimless babbling which children pursue. It appears not to be goal-directed because the untrained adult observer cannot divine a purpose other than play. Yet, it moves forward, statistically approximating adult phonetic properties more and more with age (e.g., Vihman et al. 1986, de Boysson-Bardies et al. 1992).

Such difficulties may be comparable to those faced by HABLAR's motor control architecture. Mastering speech is a hard problem because of its many articulatory and temporal degrees of freedom. HABLAR's hierarchical control architecture attempts to simplify it by introducing constraints and structure not present in a simpler articulatory skill model. However, hierarchical reinforcement learning is computationally hard despite the advantages of its divide-and-conquer strategy (Singh 1992d), even with HABLAR's added constraints and safeguards (see Section 7.4). For example, articulatory controllers'

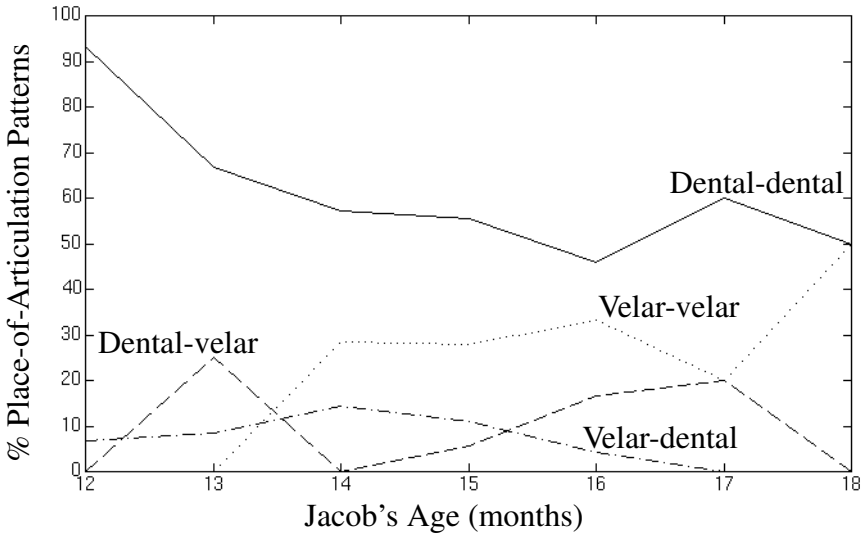


Figure 9.4 Place-of-articulation variation in a complex morpheme

Variation relative proportion of dental-velar, dental-dental, velar-velar, and velar-dental consonant sequences in Jacob's attempts to pronounce *thankyou*, never converging on any solution. Not shown are infrequent palatal sequences. Additional dimensions of variation (nasality, voicing, and vowel quality) are not shown here.

expected behaviors change over time as they learn their specialized subtasks. Thus, the phonological controller's environment (which includes unstable articulatory controllers) is not stationary. Motor learning must also overcome the combinatorial complexity of simultaneous exploration at two control levels. Indeed, there do not appear to be any principled methods for reducing the difficulties of hierarchical reinforcement learning (Singh, personal communication).

As result of these considerations, HABLAR postulates a set of heuristic bootstrapping strategies which roughly correspond to analytic strategies which children employ (see Section 7.4.7). Based on reinforcement and supervised learning simulations in a variety of compositional domains (e.g., Elman 1993, Singh 1992a-c, Tham & Prager 1994), the model is predicted to have rough going if it does not use a strategy which effectively paces development. Section 10.1.5 describes a methodology to test this conjecture.

9.2.8 Emergence of phonological awareness

Children's speech perception is organized around the syllable and encodes dynamic properties which appear to correspond to articulatory gestures (see Section 2.8.3). Their speech production is thought to be organized at first around articulatory gestures, not phoneme-sized segments (Goodell & Studdert-Kennedy 1993, Nittrouer et al. 1989, Piroli 1991; see Section 5.2), and their pronunciation errors reveal suprasegmental and even suprasyllabic patterns in which non-contiguous sounds regularly affect each other (see Section 2.4). However, children's phonological structure shows a transition from whole-word to syllabic to segmental organization as development proceeds (Macken 1979, Nittrouer et al. 1989, Vihman & Velleman 1989). Most children come to organize their speech in phoneme-sized units and become aware of such units as revealed by phoneme awareness tasks.[1] Such tasks require that phonemes be identified as individual units (e.g., selecting the non-rhyming word from among a set of words which otherwise rhyme, selecting the word whose initial consonant differs from others in a list). In normal children, phoneme

awareness precedes acquisition of reading and contributes to their ability to read. Children deficient in phoneme awareness are likely to be dyslexic (Lefly & Pennington 1995, Olson et al. 1994, Pennington et al. 1990, Wagner & Torgesen 1987). Paradoxically, children who are dyslexic or are at risk of developing dyslexia perform normally on phoneme perception (Lefly & Pennington 1995), as is generally measured by testing the ability to imitate words and nonwords with and without noise (Brady et al. 1983).

No phonemes are explicitly represented in HABLAR. The model's speech perception and production are at first organized into syllable-sized and demisyllable-sized units. For example, *duck*'s phonetic representation does not encode consonants as isolated segments, but as categories of spectral transitions (e.g., [dA], [Ak], [k^h]). Motor control may initially be organized much the same, with each articulatory controller specialized to produce only a single demisyllable or even whole consonant-vowel-consonant syllables (e.g., [dA], [k^h], [dAk]). This is a rather inefficient motor encoding, requiring as many articulatory controllers as there are demisyllables, diphthongs, and stand-alone static sounds in the language (e.g., for English we estimate about 400 categories, ignoring the effects of noncontiguous coarticulation, and not counting whole syllable types). Phonemic segmentation is more efficient than demisyllabic or syllabic segmentation, requiring many fewer articulatory controllers to generate the same set of target sounds (e.g., the number of phonemes in the language, ignoring the effects of coarticulation). Moreover, we hypothesize that it is the most efficient specialization, since a phoneme-sized unit is the least common denominator between the model's acoustic segments and its articulatory gestures and hence is the minimum possible temporal granularity of its motor control. The number of articulatory controllers is limited. Without phonemic segmentation, the system's asymptotic performance will be lower.

We hypothesize that phoneme-sized units of speech production emerge after a gradual process of differentiation as motor control and perception interact while learning a large vocabulary. Thus, *duck* would become encoded as a sequence of consonants and vowels, (e.g., /d/, /A/, /k/, and the aspirated release). This ideal segmentation into phoneme-sized articulatory schemata is portrayed in Figure 7.3. Issues of articulatory segmentation are also discussed in Section 7.3.2 and Section 7.3.4.

Phonemic segmentation should not be necessary for accurate imitation of relatively simple targets. Without phonemic segmentation, HABLAR should not be deficient in phoneme perception as measured by imitation tasks. Imitation requires only the accurate *phonetic* representation of target sounds and their articulatory interpretation by phonological and articulatory controllers, whether or not the latter are differentiated into phoneme-sized units.

However, phoneme awareness tasks are more complicated than imitation. Not only is phonemic segmentation required, but the results of such segmentation are used to solve a secondary task. Once articulatory controllers have become phonemically differentiated, HABLAR's phonemic segmentation occurs as a process of analysis-by-synthesis, in which the phonological controller activates a sequence of articulatory controllers when exposed to a target word. However, phoneme awareness requires that the results of phonemic segmentation be stored in a secondary representation and then used to solve the secondary task.

Thus, HABLAR suggests three possible hypotheses that account for deficits in phoneme awareness and possibly underlie developmental dyslexia. (1) Phonemic segmentation of motor control may not emerge. (2) The secondary representation of phonemes identified in target words may be deficient. (3) Both phonemic segmentation and secondary representation may be deficient. HABLAR also suggests two possible accounts of phonemic segmentation deficits. Temporal processing deficits (e.g., Tallal et al. 1991)

---

[1] Phonological awareness is not the same as phoneme-like perception. The former refers to the understanding that words may be segmented into phoneme-sized units. The latter refers to the perceptual assimilation of sounds to the nearest phonemic category.

may prevent accurate phonemic decomposition of sounds by disrupting acoustic segmentation of sounds. Alternatively, learning and generalization deficits may prevent the necessary differentiation of articulatory controllers into consonant and vowel specialists. For example, in order for an articulatory controller originally specialized to generate /ka/ to become a vowel specialist, it must generalize the applicability of jaw opening and tongue body lowering gestures to other subsyllabic contexts and boundary conditions. Demisyllables /ka/, /pa/, /ta/, and /ma/ contrast in the tactile feedback due to the location of articulator contact at the beginning of the syllable, but they share other properties: articulator contact, positive oral air pressure, no vibration, and relatively little motion.

9.3 Explanatory shortcomings due to HABLAR's scope or design

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. Several obvious shortcomings are beyond the scope of this work but provide opportunities for future research.

9.3.1 Coarticulation and assimilation

There are two varieties of coarticulation. Anticipatory coarticulation involves the motion of a vocal tract articulator prior to the sound for which it is required if it is not involved in the production of any intervening sounds. Thus, the lips are usually rounded during /S/ in *shoe* in anticipation of the rounded vowel. Perseveratory coarticulation involves the apparently unintentional effects of previous articulatory events on currently produced sounds. Anticipatory coarticulation seems to be planned, but perseveration is not (Whalen 1990).

Adult coarticulation and related forms of assimilation are not limited to neighboring phonemes but may jump syllable and even word boundaries (see Durand 1990). One key example is vowel harmony, in which two or more consecutive vowels share one or more distinctive features such as rounding or nasality when separated by consonants which conform to language-specific rules.

A natural account of anticipatory coarticulation and forms of multisyllabic assimilation is inconsistent with HABLAR's current design. While this is an acceptable simplification for a monosyllabic model, it is a serious shortcoming of a more general model of phonological development or adult phonology. The crucial weakness is the phonological controller's exclusive activation of a single articulatory controller at any one time. This localist articulatory representation requires a proliferation of articulatory controllers to handle allophonic variations in different coarticulatory contexts. Several possible alternatives (additional control layers, articulatory controllers with phonetic parameters) may violate computational constraints.

Phonological theory struggles with the same question. Its solution is parallelism. *The Sound Pattern of English* proposes splitting phonemes into distinctive features (Chomsky & Halle 1968). Autosegmental phonology proposes decomposing syllable structure into tiers (Goldsmith 1990).

In HABLAR, each articulatory controller crucially depends on a set of parallel agents, each of which independently manages an articulatory dimension. To accommodate anticipatory coarticulation, it may be necessary to move some of the articulatory controller's parallelism to the phonological control level. It is an open question whether hierarchical and parallel decomposition may be so thoroughly integrated, but it is certainly plausible. One possible decomposition is anatomical, dividing phonological control into seven subagents, one each for jaw, lips, tongue, hyoid, velum, glottis, and lungs. Articulatory controllers would be specialized not only by phonetic task but also by articulator. If necessary, each would distribute control of its domain among several subagents. For example, lip controllers would have sub-

agents managing height and protrusion. Tongue controllers would manage tongue body angle and displacement plus tongue tip angle and length in parallel.

9.3.2 Multisyllabic targets and utterances

HABLAR's auditory perception faithfully represents only monosyllabic stimuli. This is suitable to study the composition of many complex English single-syllable morphemes, but several barriers must be overcome before it is adequate to study multisyllabic phenomena in English or any other language. The principal change required of the articulatory system is control of pitch (i.e., fundamental frequency). To do so involves introducing another articulatory degree of freedom (glottal tension), computation of fundamental frequency by the source model, and technical changes in gesture and articulatory controller implementation.

More serious extensions are required of the model's auditory perception, including changes in the form of phonetic representation to accommodate prosodic cues (pitch and stress, changes in pitch and stress) and the relative order of syllables or demisyllables. HABLAR's simple built-in Wickelfeature encoding of relative segment order may need to give way to a better structured representation of syllable shape or a self-organized process which discovers a structured representation.

It may be appropriate to organize the overall phonetic representation hierarchically, with a morphemic level specifying prosodic structure and the relative order of syllables and a syllabic level which resembles the present model's phonetic representation. It is an open question whether such a reorganization would entail a new level of hierarchical motor control.

9.3.3 Rapid adult speech

Adult speech is often too rapid to accommodate the phonetic feedback required by the phonological controller. However, the model claims that undisturbed auditory feedback is necessary for executing speech. Slightly delayed auditory feedback severely disrupts normal speech in humans (Black 1951, Huggins 1964, Lee 1950). The worst disruption occurs for delays equal to about the average length of a syllable. Delay causes stuttering, slurring, repeated syllables, and intense frustration. However, these experiments do not eliminate feedback; they mask feedback and introduce noise. On the other hand, mid-utterance and even mid-phoneme self-corrections suggest the active monitoring of auditory feedback. In any case, adult speech may have properties of both closed-loop and open-loop control.

HABLAR's phonetic feedback is discrete and event-driven. During the interval between phonetic events, continuous proprioceptive feedback governs the articulatory control level. Thus, the interaction of phonological and articulatory control may account for some apparent off-line properties of speech.

Even so, adult speech may sometimes short-circuit explicit phonetic feedback. Likewise, adult reaction times suggest short-circuiting explicit proprioceptive feedback (e.g., Evarts 1971, Higgens & Angel 1970, Taylor & Birmingham 1948). Is it possible to do so without invalidating HABLAR's principles of closed-loop control? A more complete model with predictive components that anticipate phonetic and proprioceptive feedback achieves these goals and offers unexpected benefits.

Extending HABLAR with a predictive acoustic or phonetic component, has the effect of implementing the articulatory loop proposed by the theory of phonological working memory (Baddeley 1986, 1992). Working memory related phenomena (Baddeley 1986) suggest that articulatory dynamics are retained by the articulatory loop. Thus, we propose to insert a forward model of gestural dynamics and vocal tract acoustics between articulatory controllers and early stages of auditory perception. HABLAR's superpositional phonetic representation or its short-term memory model act as the phonological store.

As result, the model can engage in off-line babble and experimentation. It can also model observations that children with working memory deficiencies show deficits in the later stages of phonological development (Gathercole & Baddeley 1990). Certainly, the simplifying assumptions we have made are adequate for a model of early development, but the extensions suggested are attractive possibilities.

9.3.4 Articulatory limitations

HABLAR's articulatory and acoustic realism is limited in several crucial ways. Even with the addition of the source model, the articulatory synthesizer (Rubin et al. 1981) has weaknesses which limit the utterances the model can attempt, and its anatomical constraints do not always match human anatomy. For example, accurate production of liquids [l] and [r] requires a three-dimensional vocal tract model, but the synthesizer is two-dimensional. In lieu thereof, it offers control parameters which cannot be fully integrated with HABLAR's gestural model. Jaw and lip motions are more highly constrained than in humans, making it difficult to simulate labiodental [v] and [f] sounds. Furthermore, fricatives and aspiration are unnatural.

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 important factor in early phonological development. Even if the vocal tract were configurable, HABLAR does not incorporate a model of vocal tract normalization (but see Hodge 1989).

The articulatory controller's use of Q-learning limits it to discrete choices among gestural targets. Continuous variation among targets would be more natural. An alternative control strategy which retains the articulatory controller's parallelism would employ a many-actor, single-critic (e.g., Tham & Prager 1993) Adaptive Heuristic Critic architecture but would implement actors with Williams' (1992) Gaussian REINFORCE algorithm.

9.4 HABLAR and phonetic speech perception

In many ways, auditory perception is the key to HABLAR's sensorimotor integration. It defines the representational form and contents of utterances. Its phonetic representation defines dimensions of phonetic similarity and acts as the foundation for generalization and compositionality. Acoustic segmentation provides the event-driven timing needed for hierarchical motor control. Acoustic segmentation and phonetic categorization define the abstract state space employed by hierarchical motor control. Reinforcement signals which guide motor learning depend on phonetic representations of the target utterance and auditory feedback.

Despite this central role, HABLAR is designed to test patterns of speech production, not to test auditory or phonetic perception. Still, HABLAR offers explanations of perceptual phenomena. It models the categorical perception of speech which develops during infancy (Section 6.4.5). With appropriate extensions and methodology, we believe it will explain the emergence of phonological awareness (Section 9.2.8). In this section, we explore how it might explain other high-level perceptual phenomena, and in particular how its account compares with the motor theory of speech (e.g., Liberman & Mattingly 1985) and the theory of direct perception (e.g., Fowler & Rosenblum 1991).

9.4.1 Motor and direct perception theories of speech perception

The basic question addressed by motor and direct perception theories of speech perception is the problem of invariant perceptual cues. It is observed that the same static acoustic cue often gives rise to dif-

ferent percepts in different contexts, but different cues often give rise to the same percept in different contexts. Attempts to identify invariant static cues which underlie speech perception have failed; dynamic acoustic cues seem to be essential instead (see Section 2.8.3). Motor and direct perception theories each argue that what is invariant are linguistically organized phonetic gestures, not instrumentally observable acoustic events and not individually observable motor events. They contend that gestures are perceived directly, not mediated by prior stages of acoustic and phonetic analysis proposed by process-oriented cognitive models (e.g., Sawusch 1986). It is here, however, that the motor theory and the theory of direct perception diverge.

Motor theory (Liberman & Mattingly 1985) contends that both speech perception and speech production draw on the same set of invariants, explaining the articulatory characteristics of speech perception. For example, it claims that there is some abstract notion of "lip rounding" that underlies both the articulatory program to accomplish lip rounding and the perception of lip rounding. Thus, the link between perception and production is not learned but is some innate neural structure. It is claimed to be a privileged speech perception module which enables the direct perception of articulatory gestures without intervening auditory analysis (Liberman & Mattingly 1989).

On the other hand, the theory of direct perception (Fowler & Rosenblum 1991) contends that the perception of distal events is not mediated by motor systems or any component shared with the motor control of speech. It rejects the notion of a privileged, special, or closed module. Instead, it takes the Gibsonian stance that all perception is direct (Gibson 1966, 1979). Percepts are meaningful distal events. Proximal signals (e.g., context-sensitive acoustic cues) are merely the medium of transmission. Phonetic gestures are the distal events which are recovered by speech perception from proximal auditory stimulation. Furthermore, direct perception is able to unravel overlapping gestures whose acoustic signals are combined (Fowler & Smith 1986, Fowler & Rosenblum 1991). For example, pitch conveys both linguistic and pragmatic information. Despite the many determinants of absolute pitch (e.g., intonation gestures which stretch the vocal folds, respiration, tongue motion which incidentally stretches the vocal folds), the listener unravels and attributes variations on pitch to each source, accurately decoding linguistic (vowel height) and pragmatic (intonation) information.

Neither theory offers further computational details of how direct perception occurs, of what information present in the speech signal is used to recover the distal stimulus, or of how the distal stimulus is recovered. In his commentary on the Fowler & Rosenblum paper, MacNeilage (1991) asserts that neither motor theorists nor direct realists have adequately defined "phonetic gesture", and that neither has explained how gestures might underlie perception (since production always lags behind perception developmentally).

According to MacNeilage the direct perception assumption that there is no distinction between surface (auditory) and underlying (phonetic) representations of speech sounds is unjustified. In this regard, Sawusch (1986) reviews psychophysical evidence clearly showing perceptual effects due to distinct auditory processes (such as selective adaptation) and phonetic processes (Ainsworth 1977, Sawusch & Jusczyk 1981, Sawusch & Nusbaum 1983).

9.4.2 Grounding the notion of direct perception

What computational process might ground the notion of direct perception, explain the articulatory character of phonetic perception, yet acknowledge effects due to auditory processing? The resolution of this dilemma is to observe that it is possible build a statistical mixture model of the environment which learns its (possibly overlapping) invariant features *by modeling the processes which generate them* (Nowlan 1991a,b; see Section 6.2.5).

Assume that properties of phonetic events may be characterized not as one uniform distribution, but as clusters of phonetically similar events, which can be described as a mixture of many distributions. A statistical mixture model then assumes that *each cluster* may be modeled by a separate parametric statistical model, and that the whole population of phonetic events may be modeled as a mixture of such parametric statistical models. More generally, since we are considering dynamic speech events, the population may be modeled as being generated by a collection of parametric stochastic processes. In this paradigm, recognition of a phonetic event consists in determining which stochastic process is most likely to have generated the event.

Thus, a statistical mixture model for the recognition of continuous speech consists of a set of stochastic processes whose parameters may be adjusted to simulate (and perhaps generate) the set of linguistically organized phonetic gestures in the environment, a procedure for adjusting stochastic process parameters, and a model which estimates the likelihood with which each stochastic process generates an observed sound. Such a model can employ an approach like the EM algorithm (Dempster et al. 1977) to learn the linguistic environment's characteristics. For each observed sound, the algorithm iteratively adjusts the parameters of each stochastic process to maximize its likelihood of generating the sound, proportional to the likelihood that it generated it in the first place.

This approach precisely describes HABLAR's motor control apparatus and how it models the linguistic environment (see Section 9.1.5). The set of stochastic processes which simulate linguistically organized phonetic gestures are HABLAR's articulatory controllers. The model which estimates the likelihood with which each stochastic process generates an observed sound is HABLAR's phonological controller. The procedure for adjusting stochastic process parameters is the reinforcement learning method employed by the articulatory controllers. Likewise, the procedure for adjusting the likelihood model is the reinforcement learning method used by the phonological controller. Sound recognition consists of the phonological controller's identifying the sequence of articulatory controllers which is most likely to generate the observed sound.

This account acknowledges the primacy of functionally coordinated sets of gestures and provides a proximal method for estimating the distal gesture from the speech signal. Thus, it grounds the direct perception of distal phonetic events. Like the motor theory, it suggests that the link between production and perception is common process, though it makes explicit the computational nature of the process. When extending HABLAR as suggested in Section 9.3.1 to better account for coarticulatory patterns of speech production, the model suggests a mechanism for structural credit assignment among overlapping gestures.

HABLAR thus suggests an implementation and unification of the motor and direct perception theories, but it is not equivalent to either theory. The phonological controller observes sounds, but it does not match them with actual gestures. It matches them with an abstract model most likely to have generated each sound. Although nominally a motor component, the phonological controller's role is phonetic and phonological. Besides, it is probably impossible to unravel its perception and production responsibilities. The phonological controller is specialized in its tasks, but it clearly does not meet the criteria for a closed or privileged module (Fodor 1983).

HABLAR does not short-circuit acoustic, auditory, or phonetic analysis of speech signals. It acknowledges the role of each processing stage in speech perception. Moreover, HABLAR's recoding of phonetic inputs in abstract articulatory terms is an example of the re-representation process proposed by Karmiloff-Smith (1992) as underlying higher levels of cognition. We have not yet considered how the output of this process might be stored or employed in other phonological processes. Also unexplained is why perception comes to attend the analysis of the phonological controller instead of auditory or phonetic representations generated by auditory perception. That is a question beyond the scope of this research.

Chapter 10 Conclusions and Future Directions

The goal of this research is a computational model of the sensorimotor foundations of phonology which explains key characteristics of phonological development. We have accomplished part of the goal; the model is built. We have demonstrated some of its capabilities. We have provided qualitative accounts of the data. However, we have not yet conducted definitive quantitative simulations. An assessment of its predictions also suggests a number of extensions and modifications.

The purpose of this chapter is to outline possible future directions in this research. We propose a set of simulations to test HABLAR's fidelity and scientific validity and summarize extensions of the model which we intend to explore. Finally, we conclude the thesis with a summary of the work's contributions.

10.1 Design for an empirical test of HABLAR's fidelity

With few exceptions, the key generalizations of psycholinguistic studies of phonological development are qualitative. Most are based on longitudinal studies of individual children (e.g., Menn 1971, 1976, Waterson 1971, Smith 1973, Macken 1979). More recently some quantitative assessments using larger samples and cross-linguistic controls have appeared (e.g., de Boysson-Bardies et al. 1989, de Boysson-Bardies and Vihman 1991, de Boysson-Bardies et al. 1992, Kuhl et al. 1992, Vihman et al. 1986, Werker et al. 1981). Even these studies have been limited to simple cross-linguistic comparisons of phonetic perception or the phonetic properties of babble and early speech. Children's developmental paths are so idiosyncratic (Vihman 1993) that a common basis on which to compare them is elusive, especially beyond the earliest stages. Controlled experiments within a linguistic community are not possible without observing and manipulating parent-child interactions. Experimental manipulations are usually fortuitous (Menn 1971) or of limited applicability (e.g., Berko 1958).

Thus, to test the fidelity of the model's simulation of human phonological development, we compare its qualitative behavior with that observed for children in a number of key areas. In some cases we may use experimental manipulations or directly inspect internal representations in ways not possible with real children. The most direct experimental control common to all simulations is HABLAR's linguistic environment. Parental vocabulary is selected from a synthetically generated corpus 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, (4) represent a reasonable distribution of sound patterns from which the compositional structure of one-syllable utterances may be induced, and (5) conform to a chosen experimental design. For some experimental manipulations, we may compile a vocabulary from actual samples of child-directed speech, drawing from the CHILDES database (MacWhinney 1991).

The basic observational method common to most experiments will be phonetic transcription of speech samples with specific structural, longitudinal, and statistical analyses. Some if not most transcriptions will be automatic, using formant frequencies to classify vowels, a neural network classifier trained to recognize parental syllables, and a set of rules to analyze the syllabic structure of utterances based on the model's own acoustic segmentation algorithm. The validity of automated transcriptions will be tested against human observers trained in phonetic transcription, and if necessary it will be adapted and adjusted such that it achieves cross-transcription agreement rates comparable to typical studies of child phonology (e.g., Oller et al. 1976, Vihman et al. 1986). Unsystematic validation of the auditory perception module's segmentation and classification of its own speech is encouraging in this regard.

The key properties we shall test are (1) changes in the phonetic characteristics of babble, (2) patterns of sound substitutions and deletions, (3) the systematicity of pronunciation errors, (4) overgeneralization, (5) pronunciation difficulties due to inefficient learning strategies, and (6) the emergence of phoneme-sized phonological units and phonological awareness. For each of these we shall explain the method used to test the model's behavior.

10.1.1 The phonetic characteristics of babble

The simplest test of babble's increasing approximation of adult phonetics is a comparison of phonetic statistics for model and parental corpus. Lacking any control, this method will be reserved for initial informal tests of the model's behavior. A formal test will introduce several model instances divided into two groups. Each group will be exposed to parental corpora with statistically different phonetic characteristics (e.g., distribution of vowels, stops, nasals, places of articulation). If the model behaves like children, at first the mean phonetic properties of the two groups should be similar and their variances high; then their means should diverge, more closely approximating parental phonetics, and their variances should fall.

10.1.2 Sound substitution and deletion

To test simple substitutions and deletions requires a statistical compilation of the model's segmental substitutions or deletions at various stages of development. However, direct comparisons between human and model behaviors will be misleading because the model's anatomical constraints do not match human anatomy. Still, if the model is correct, the types of substitutions HABLAR makes should qualitatively resemble those by children.

10.1.3 Context sensitive systematicity of pronunciation errors

The context sensitivity and systematicity of pronunciation errors will be captured by conditional statistics or by one of several methods used by linguists: rewrite rules which describe the errors and their contexts (e.g., Menn 1971, Smith 1973), or a set of canonical forms or prosodic structures which describe contextual regularities (Macken 1979, Menn 1976, Waterson 1971). Given the idiosyncratic nature of such context sensitive errors among children, and without any known statistical assessment of their systematicity, we can only qualitatively compare HABLAR's behavior with human behavior. However, to experimentally confirm its behavior we will introduce new words which conform to the suspected patterns and observe the model's pronunciations.

We conjecture that some context sensitive error patterns (e.g., consonant harmony) are partly due to the effect of idiosyncratic sound preferences and developmental history which bias new sound patterns. To test this hypothesis, we will experimentally manipulate developmental path and vocabulary introduction for some model instances.

10.1.4 Overgeneralization and inertia

As a test of overgeneralization and inertia, we introduce new words which differ from the model's repertory along specific dimensions and observe the model's pronunciation. We also manipulate the schedule with which words are introduced.

10.1.5 Articulatory precision and bootstrapping strategies

We hypothesize that "analytic" and "gestalt" speaking styles and apparent differences in articulatory precision observed by Peters (1977) are a result of differences in the rate with which the complexity of target vocabulary grows, that is, differences in children's internal selective strategies. To test this supposition and to test different selective strategies, we shall compare the behavior of two groups of model instances, one group exposed to a lexicon which grows in phonological complexity gradually, and the other group exposed to the same lexicon in random order. Additional experimental manipulations will test the effect of differences in (1) error tolerance, and (2) specific lexical and phonetic preferences.

10.1.6 Emergence of phonological segmentation and awareness

To test HABLAR's evolving phonemic segmentation, we can directly inspect the behavior of the phonological controller to see how it parcels the pronunciation of complex utterances among several articulatory controllers. To induce deficits in phonemic segmentation caused by temporal processing deficits, we shall experiment with delays introduced into auditory perception. To induce deficits in phonemic segmentation caused by learning deficits, we will manipulate connectivity, memory constraints, and factors influencing generalization. To test phoneme awareness, we will augment the model with a secondary representation of the sequence of phonological controller actions and with an additional component which is trained to perform one or more phoneme awareness tasks. We will test its ability to perform phoneme awareness tasks under a range of phonemic segmentation conditions, comparing its behavior with results obtained from the literature.

10.2 Candidate model extensions and applications

In Chapter 9, we identify a number of HABLAR's weaknesses. Chief among these is its failure to adequately account for anticipatory coarticulation observed in speech — the motion of an articulator which anticipates a future sound and colors the phonetic properties of intervening sounds. Exclusive activation of one articulatory controller at a time makes it difficult to simulate normal patterns of coarticulation. A scheme in which several articulatory controllers might be active at any one time will enable the parallel execution of several functionally coordinated articulatory gestures. Introducing parallelism to the phonological control level should also augment HABLAR's account of direct perception, enabling the model to unravel the acoustic signals of overlapping gestures. It may be that some developmental phenomena depend on the child's inability to accurately unravel these sounds. Consonant harmony may owe some of its systematicity to the way in which a distributed phonological control system assigns credit for component sounds among overlapping articulatory programs. Thus, this extension has a particularly high priority.

Adding the capacity for multisyllabic speech perception and production to HABLAR is essential for any cross-linguistic validation of the model. As suggested in Chapter 9, perceptual representations and articulatory capabilities must be augmented for this to occur.

Currently speech in HABLAR is exclusively synthetic. This is a function of practical considerations and the controlled linguistic environment necessary to test the model's validity (Section 10.1). However, an exhaustive test of HABLAR's validity should include an evaluation of how well its auditory perception module can segment and categorize natural speech. Sawusch (personal communication) argues that dynamic spectral properties are not as well behaved in natural speech as in synthetic speech. However, analysis of adult and children's speech using a regression model of consonant-vowel coarticulation (Sussman et al. 1994) and acoustic analysis of children's consonant-vowel gestures (Piroli 1991) suggest well-defined dynamic acoustic cues. Application of HABLAR's auditory perception to real speech requires at

least the addition of a signal processing front end (e.g., Seneff 1988). If the evaluation is successful, its application to automatic speech recognition is a possibility.

10.3 Contributions

HABLAR is distinguished in several respects. It models human performance over a broad range of high-level cognitive tasks. It models an integrated system situated in a realistic though controlled environment. As a large scale sensorimotor and developmental model of complex human behavior, it faces certain disadvantages. Most serious is the difficulty of testing and validating its claims. Even so, it makes assumptions usually missing from most limited-scope, single-task models quite explicit by simulating, as it does, an entire sensorimotor system.

10.3.1 Technical features

HABLAR features auditory and articulatory systems and a cognitive architecture which bridges the two. Several components make specific technical contributions. Some offer new ways of thinking about speech perception and production by applying known computational methods to old psychological and linguistic dilemmas.

Auditory perception contributes a method to segment and categorize speech sounds based exclusively on acoustic cues. It employs soft competitive learning (Nowlan 1991a,b), a self-organized process, to learn categories of vowels and consonant-vowel demisyllables which mimic the categorical perception which emerges in late infancy. Furthermore, auditory perception is adapted to the needs of motor control and plays a central role in sensorimotor integration.

Hierarchical control integrates continuous, fine-grained control and event-driven compositional control. It employs Q-learning (Watkins 1989) to link perceptual and articulatory events. The model's auditory perception is specialized to segment and categorize acoustic feedback into discrete phonetic events which closely correspond to discrete functionally coordinated gestures learned by articulatory controllers. HABLAR need not solve the hard problem of relating continuous speech sound and continuous vocal tract motion. It learns the correspondence between one discrete sequence of events and another.

Hierarchical control also plays a perceptual role. By performing a kind of analysis-by-synthesis of speech stimuli, it can account for the emergence of phonological awareness and may be useful in understanding why and when deficiencies in phonological awareness occur. It also provides a computational grounding for the direct perception of speech gestures, possibly implementing and unifying the motor theory of speech (Liberman & Mattingly 1985) and the theory of direct perception (Fowler & Rosenblum 1991).

HABLAR introduces and applies a parallel reinforcement learning architecture, a parallel version of Q-learning, to learn articulatory control of the vocal tract's many degrees of freedom.

10.3.2 Systemic constraints on neural network hypotheses

Language's complexity and the relative ease with which most children acquire language suggests built-in knowledge (e.g., Chomsky 1965). What form does that knowledge take? Certainly part may be encoded in the intricacies of human neural architecture. Yet, the architecture must be flexible enough to accommodate any language, physical environment, and culture.

Is it possible to merely discover regularities in the environment? Unfortunately, neural networks are not immune to the laws of statistics (Smolensky 1994). Without constraints on a network's hypothesis space, it is unlikely to learn the right generalizations or discover the appropriate regularities in the environ-

ment (Geman et al. 1992). Furthermore, constraints on what is learnable and the architecture of linguistically relevant cognitive machinery may shape the regularities observed in language. Are we thus forced to explicitly introduce prior knowledge?

Real neural computation occurs in a large systemic context. Anatomy, motor dynamics, drives, child-parent interactions, and — in the case of phonological development — speech articulators, vocal tract acoustics, and auditory feature detectors specialized for speech, may govern development no less than the prior knowledge built into the system's overall neural architecture and individual neural components. To find the most parsimonious account, one must explore this larger context. Knowledge may be stored in the world (Ballard 1991), not just in the language acquisition device. HABLAR takes this approach, modeling emergent linguistic behavior with minimal linguistic assumptions. Its hypothesis space is constrained by its architecture, the form of its internal representations, interactions among its components, and interactions with its linguistic environment.

Complexity has its costs — and its payoffs.

10.3.3 A new testbed and analytic framework

HABLAR offers a useful computational testbed for evaluating theories of phonological development and broader questions of language acquisition. It also provides a new framework within which to formulate and assess hypotheses about phonological development. Our task now is to experimentally demonstrate its scientific validity. That work is already underway.

Bibliography

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

Ackley, D.H. & Littman, M.L. (1990). Generalization and scaling in reinforcement learning. In D.S. Touretzky (Ed.), Advances in *Neural Information Processing Systems, 2:* 550-557. San Mateo, CA: Morgan Kaufmann.

Ainsworth, W.A. (1977). Mechanisms of selective feature adaptation. *Perception and Psychophysics, 21:* 365-370.

Anderson, C.W. (1987). Strategy learning with multilayer connectionist representations (Technical report TR87-509.3). Waltham, MA: GTE Laboratories.

Archangeli, D. (1984). *Underspecification in Yawelmani phonology and morphology.* Ph.D. Thesis. Cambridge, MA: MIT, Department of Linguistics and Philosophy.

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

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

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

Bailly, G., Laboissiere, R., and J.L. Schwartz (1991). Formant trajectories as audible gestures: an alternative for speech synthesis. *Journal of Phonetics, 19:* 9-23.

Baird, L.C. and Klopf, A.H. (1993). Reinforcement learning with high-dimensional, continuous actions. Technical Report WL-TR-93-1147, Wright-Patterson Air Force Base, Ohio.

Ballard, D.H. (1991). Animate vision. *Artificial Intelligence, 48:* 57-86.

Ballard, D.H. and Brown, C.M. (1992). Principles of Animate Vision. *CVGIP: Image Understanding, 56:* 3-21.

Barto, A.G. (1985). Learning by statistical cooperation of self-interested neuron-like computing elements. *Human Neurobiology, 4:* 229-256.

Barto, A.G., Sutton, R.S., & Anderson, C.W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics, 13:* 834-846.

Barto, A.G., Sutton, R.S., & Watkins, C.J.C.H. (1990). Learning and sequential decision making. In M. Gabriel & J. Moore (eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks,* Chapter 13, pp. 539-602. Cambridge, MA: MIT Press.

Barton, D. (1976). *The role of perception in the acquisition of phonology.* Ph.D. Dissertation. University of London.

Barton, D. (1980). Phonemic perception in children. In G.H. Yeni-Komshian, J.F. Kavanagh, and C.A. Ferguson (Eds.), *Child Phonology,* Volume 2: Perception, pp. 97-116. New York: Academic Press.

Bellman, R. (1957). *Dynamic programming.* Princeton, NJ: Princeton 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.

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

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

Bertoncini, J. and Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development, 4:* 247-260.

Bertsekas, D.P. (1976). *Dynamic programming and stochastic control.* New York: Academic Press.

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

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

Blumstein, S.E., Isaacs, E., & Mertus, J.(1982). The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America, 72:* 43-50.

Boyan, J.A. (1992). *Modular neural networks for learning context-dependent game strategies.* Master's Thesis. University of Cambridge.

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.

Brady, S., Shankweiler, D., & Mann, V.A. (1983). Speech perception and memory coding in relation to reading ability. *Journal of Experimental Child Psychology*, 35: 345-367.

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

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

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

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., and Gerstman, L.J. (1952). Some experiments on the perception of synthetic speech. *Journal of the Acoustical Society of America, 24:* 597-606.

Creaghead, Nancy A., Newman, P.W., Secord, W.A. 1989. *Assessment and Remediation of Articulatory and Phonological Disorders.* 2nd Ed. NY: Macmillan.

Crites, R.H. (1994). Multi-agent reinforcement learning. Thesis proposal. University of Massachusetts: Department of Computer Science.

Crystal, D. (1985). *A dictionary of linguistics and phonetics.* 2nd Edition. Oxford, UK: Basil Blackwell Ltd.

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.

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.

Delattre, P.C., Liberman, A.M., and Cooper, F.S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27:* 769-773.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B, 39:* 1-38.

Doya, K. and Sejnowski, T.J. (1995). A novel reinforcement model of birdsong vocalization learning. To appear in *Neural Information Processing Systems, 7*.

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

Edwards, M.L. (1974). Perception and production in child phonology: the testing of four hypotheses. *Journal of Child Language, 1:* 205-219.

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48:* 71-99.

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.

Flanagan, J.L., K. Ishizaka, and K.L. Shipley (1975). Synthesis of speech from a dynamic model of the vocal cords and vocal tract. *The Bell System Technical Journal, 54* (3): 485-506.

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

Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14:* 3-28.

Fowler, C.A. and Rosenblum, L.D. (1991). The perception of phonetic gestures. In I.G. Mattingly and M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception,* pp. 33-59. Hillsdale, NJ: Lawrence Erlbaum Associates.

Fowler, C.A. and Smith, M.R. (1986). Speech perception as 'vector analysis': an approach to the problems of invariance and segmentation. In J.S. Perkell and D.H. Klatt (Eds.), *Invariance and Variability in Speech Processes,* pp. 123-139. Hillsdale, NJ: Lawrence Erlbaum.

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

Gasser, M. (1992). Learning distributed representations for syllables. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society,* pp. 396-401. Hillsdale, NJ: Erlbaum.

Gasser, M. (1993). Learning words in time: towards a modular connectionist account of the acquisition of receptive morphology. Technical Report 384. Bloomington, IN: Indiana University Dept. of Computer Science, June 10, 1993.

Gasser, M. & Lee, C.-D. (1990). A short-term memory architecture for the learning of morphophonemic rules. In R.P. Lippmann, J.E. Moody, and D.S. Touretsky (Eds.), *Advances in Neural Information Processing Systems, 3:* 605-611. San Mateo, CA: Morgan Kaufmann Publishers

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.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4:* 1-58.

Gerken, L.A. (1994). Child phonology: past research, present questions, future directions. In M.A.Gernsbacher (Ed.), *Handbook of Psycholinguistics,* pp. 781-820. San Diego: Academic Press.

Gibson, J.J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Gibson, J.J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Goldsmith, J.A. (1990). *Autosegmental and metrical phonology.* Oxford: Basil Blackwell.

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

Goodell, E.W. and Studdert-Kennedy, M. (1993). Acoustic evidence for the development of gestural coordination in the speech of 2-year-olds: a longitudinal study. *Journal of Speech and Hearing Research, 36:* 707-727

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.

Guenther, F.H. (1994). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, in press.

Gullapalli, V. (1993). Learning control under extreme uncertainty. In S.J. Hanson, J.D. Cowan and C.L. Giles (Eds.), *Neural Information Processing Systems, 5:* 271-278.

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.

Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., and Jordan, M.I. (1992). Forward dynamics modeling of speech motor control using physiological data. In J.E. Moody, S.J. Hanson, and R.P. Lippmann (Eds.), *Neural Information Processing Systems, 4:* 191-198.

Hirayama, M., Vatikiotis-Bateson, E., Honda, K., Koike, Y. and Kawato, M. (1993). Physiologically based speech synthesis. In S.J. Hanson, J.D. Cowan and C.L. Giles (Eds.), *Neural Information Processing Systems, 5:* 658-665.

Hirayama, M., Vatikiotis-Bateson, E., and Kawato, M. (1994). Inverse dynamics of speech motor control. In J.D. Cowan, G. Tesauro, and J. Alspector (Eds.), *Neural Information Processing Systems, 6:* 1043-1050.

Hodge, M. (1989). *Dynamic spectral-temporal characteristics of children's speech. Implications for a model of speech skill development.* Ph.D. Dissertation. University of Wisconsin, Madison.

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.

Jakobson, R. (1941/1968). *Child Language, Aphasia and Linguistic Universals.* The Hague: Mouton.

Jordan, M.I. (1986). Serial order: a parallel distributed processing approach. Technical Report ICS-8604, May 1986. Institute for Cognitive Science, University of California, San Diego.

Jordan, M.I. (1990). Motor learning and the degrees of freedom problem. In M. Jeannerod (Ed.), *Attention and Performance XIII: Motor Representation and Control,* pp. 796-836. Hillsdale, NJ: Lawrence Erlbaum.

Jordan, M.I. and Jacobs, R.A. (1990). Learning to control an unstable system with forward modeling. In D.S. Touretzky (Ed.), *Advances in Neural Information Processing Systems, 2:* 324-331. San Mateo, CA: Morgan Kaufmann.

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.

Jusczyk, P.W. (1993). From general to language-specific capacities: the WRAPSA Model of how speech perception develops. *Journal of Phonetics, 21:* 3-28.

Jusczyk, P.W. and Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology, 23:* 648-654.

Jusczyk, P.W., Jusczyk, A.M., Kennedy, L.J., Schomberg, T. and Koenig, N. (1995). Young infants' retention of information about bisyllabic utterances. *Journal of Experimental Psychology: Human Perception and Performance*. In press.

Jusczyk, P.W., Luce, P.A., Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33:*630-645.

Kaelbling, L.P. (1993). Hierarchical learning in stochastic domains: preliminary results. In *Machine Learning: Proceedings of the Tenth International Conference,* pp. 167-173. San Mateo, CA: Morgan Kaufmann.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science.* Cambridge, MA: MIT Press.

Kelso, J.A.S., Holt, K.G., Kugler, P.N. & Turvey, M.T. (1980). On the concept of coordinative structures as dissipative structures: II. Empirical lines of convergence. In G.E. Stelmach and J. Requin (Eds.), *Tutorials in Motor Behavior,* pp. 49-70. Amsterdam: North-Holland Publishing Company.

Kelso, J.A.S., Saltzman, E.L. & Tuller, B. (1986). The dynamical perspective on speech production: data and theory. *Journal of Phonetics, 14:* 29-59.

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. (1992). The biology of phonological development. In C.A. Ferguson, L. Menn, and C. Stoel-Gammon (eds.), *Phonological Development: Models, Research, Implications,* p. 65-90. Parkton, MD: York Press.

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

Kewley-Port, D., Pisoni, D.B., Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *Journal of the Acoustical Society of America, 73:* 1779-1793.

Kiparsky, P. and Menn, L. (1977). On the acquisition of phonology. In J. Macnamara (Ed.), *Language learning and thought.* New York: Academic Press.

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.

Ladefoged, Peter (1982). *A Course in Phonetics.* 2nd Ed. Harcourt Brace Jovanovich, New York.

Lahiri, A., Gewirth, L., and Blumstein, S.E. (1984). A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. *Journal of Acoustical Society of America, 73:* 1003-1010.

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

Lee, C.-D. and Gasser, M. (1992). Where do underlying representations come from: a connectionist approach to the acquisition of phonological rules. In J. Dinsmore (Ed.), *The symbolic and connectionist paradigms: closing the gap,* pp. 179-207. Hillsdale, NJ: Lawrence Erlbaum.

Lee, K.-F. (1990). Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 38:* 599-609.

Lefly, D.L. and Pennington, B.F. (1995). Longitudinal study of children at high family risk for dyslexia: the first two years. In M.L. Rice (Ed.), *Toward a genetics of language*. Hillsdale, NJ: Erlbaum. In press.

Levelt, C.C. (1994). *On the acquisition of place.* Ph.D. Thesis. Leiden: HIL Dissertations in Linguistics.

Levitzky, M.G. (1991). *Pulmonary Physiology,* 3rd Ed. New York: McGraw-Hill.

Liberman, A.M., and Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition, 21:* 1-36.

Liberman, A.M., and Mattingly, I.G. (1989). A specialization for speech perception. *Science, 243:* 489-494.

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

Lin, L.-J. (1993a). Scaling up reinforcement learning for robot control. In *Machine Learning: Proceedings of the Tenth International Conference,* pp. 182-189. San Mateo, CA: Morgan Kaufmann.

Lin, L.-J. (1993b). *Reinforcement learning for robots using neural networks.* Ph.D. Thesis. Pittsburgh, PA: Carnegie Mellon University.

Lin, L.-J. & Mitchell, T.M. (1992). Memory approaches to reinforcement learning in non-Markovian domains. Technical Report CMU-CS-92-138. Pittsburgh, PA: Carnegie Mellon University, School of Computer Science.

Lindblom, B. (1991). The status of phonetic gestures. In I.G. Mattingly and M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception*, pp. 7-31. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lindblom, B. (1992). Phonological units as adaptive emergents of lexical development. In C.A. Ferguson, L. Menn, and C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications*, pp. 131-163. Parkton, MD: York Press.

Lindblom, B., Lubker, J., and Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation, *Journal of Phonetics, 7:* 147-161

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

Ling, C.X. and Marinov, M. (1993). Answering the connectionist challenge: a symbolic model of learning the past tenses of English verbs. *Cognition, 49:* 235-290.

Locke, John L. (1983). *Phonological acquisition and change*. New York: Academic Press.

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

Macken, M.A. and Barton, D. (1979). The acquisition of the voicing contrast in English: a study of voice onset time in word-initial stop consonants. *Journal of Child Language, 7:* 41-74.

MacNeilage, P.F. (1980). The control of speech production. In G.H. Yeni-Komshian, J.F. Kavanagh, and C.A. Ferguson (Eds.), *Child Phonology, Volume 1: Production,* pp. 9-21. New York: Academic Press.

MacNeilage, P.F. (1991). Comment: The gesture as a unit in speech perception theories. In I.G. Mattingly and M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception,* pp. 61-67. Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (1978). *The acquisition of morphophonology.* Monographs of the Society for Research in Child Development. #174, vol. 43.

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.

Mahadevan, S. and Connell, J. (1991). Scaling reinforcement learning to robotics by exploiting the subsumption architecture. In *Machine Learning: Proceedings of the Eighth International Conference,* pp. 328-332. San Mateo, CA: Morgan Kaufmann.

Markey, K. (1994a). 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.

Markey, K.L. (1994b). Acoustic-based syllabic representation and articulatory gesture detection: prerequisites for early childhood phonetic and articulatory development. In A.Ram & K. Eiselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society,* pp. 595-600, Hillsdale, NJ: Erlbaum.

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

Markey, K.L., Dodier, R., and Mozer, M.C. (in preparation). Parallel reinforcement-based learning of multi-dimensional control tasks.

Markey, K.L. and Mozer, M.C. (1992). Comparison of reinforcement algorithms on discrete functions: learnability, time complexity, and scaling. *Proceedings of the International Joint Conference on Neural Networks* (Baltimore), I-853-859. Piscataway, NJ: IEEE.

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

Mehler, J., Dupoux, E. and Segui, J. (1990). Constraining models of lexical access: the onset of word recognition. In G.T.M. Altmann (Ed.), *Cognitive models of speech processing.* Hillsdale, NJ: Lawrence Erlbaum

Mehler, J., Jusczyk, P.W., Lambertz, G., Halsted, N., Bertoncini, J. and Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29:* 143-178.

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. (1979). Transition and variation in child phonology: modeling a developing system. Paper presented for the International Congress of Phonetic Sciences, Copenhagen.

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,* pp. 211-247. Parkton, MD: York Press.

Menn, L., Markey, K., Mozer, M. and Lewis, C. (1993). 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,* pp. 421-433. 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.

Mines, A.H. (1993). *Respiratory Physiology,* 3rd Ed. New York: Raven Press.

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 S.J. Hanson et al. (Eds.), *Neural Information Processing Systems, 5:* 969-976.

Moody, J. and Tresp, V. (1994). A trivial but fast reinforcement controller. *Neural Computation*, 6.

Munro, P. (1987). A dual back-propagation scheme for scalar reward learning. In *Program of the Ninth Annual Conference of the Cognitive Science Society,* 165-176. Hillsdale, NJ: Erlbaum.

Narendra, K.S. and Thathachar, M.A.L. (1989). *Learning Automata: An Introduction.* Englewood Cliffs, NJ: Prentice Hall.

Nash, J. (1950). Equlibrium points in *N*-person games. *Proceedings of the National Academy of Sciences, 36:* 48-49.

Neisser, U. (1967). *Cognitive Psychology.* New York: Appleton-Century-Crofts.

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

Nittrouer, S. and Studdert-Kennedy, M. (1987). The role of coarticulatory effects in the perception of fricatives by children and adults. *Journal of Speech and Hearing Research, 30:* 319-329.

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

Nossair, Z.B. and Zahorian, S.A. (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America, 89:* 2978-2991.

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

Nowlan, S.J. (1991b). *Soft competitive adaptation: neural network learning algorithms based on fitting statistical measures.* Ph.D. Dissertation. Technical Report CMU-CS-91-126. Carnegie Mellon University: School of Computer Science.

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. and MacNeilage, P.F. (1983). Development of speech production: Perspectives from natural and perturbed speech. In P.F. MacNeilage (Ed.), *The production of speech*, pp. 91-108. New York: Springer-Verlag.

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

Pennington, B.F., Van Orden, G.C., Smith, S.D., Green, P.A. and Haith, M.M. (1990). Phonological processing skills and deficits in adult dyslexics. *Child Development, 61:* 1753-1778.

Peters, A.M. (1977). Language learning strategies: does the whole equal the sum of the parts? *Language, 53:* 560-573.

Piaget, J. (1952/1963). *The origins of intelligence in children.* New York: W.W. Norton & Company.

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

Piroli, James R. (1991). *An acoustic typology of infant protosyllables.* Ph.D. Thesis, May, 1991, Department of Communication Disorders, Louisiana State University.

Plunkett, K. and Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron. *Cognition, 39:* 43-102.

Polit, A. & Bizzi, E. (1978). Processes controlling arm movements in monkeys. *Science,* 201: 1235-1237.

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

Prescott, T. & Mayhew, J. (1992). Obstacle avoidance through reinforcement learning. In J.E. Moody et. al. (eds.), *Advances in Neural Information Processing Systems, 4:* 523-530. San Mateo, CA: Morgan Kaufmann.

Priestly, T.M.S (1977). One idiosyncratic strategy in the acquisition of phonology. *Journal of Child Language, 4:* 45-66.

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

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel distributed processing: Explorations in the Microstructure of Cognition,* vol. 1, pp. 318-362. Cambridge, MA: MIT Press.

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.

Saltzman, E.L. and Munhall, K.G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1 (4):* 333-382.

Sander, E. (1972). When are speech sounds learned? *Journal of Speech and Hearing Disorders, 37:* 55-63.

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

Sawusch, J.R. (1986). Auditory & phonetic coding of speech. In E.C. Schwab and H.C. Nusbaum, *Pattern recognition by humans & machines, Vol. 1: Speech perception,* pp. 51-88. Orlando, FL: Academic Press.

Sawusch, J.R. & Jusczyk, P.W. (1981). Adaptation and contrast in the perception of voicing. *Journal of Experimental Psychology: Human Perception and Performance, 7:* 408-421.

Sawusch, J.R. & Nusbaum, H.C. (1983). Auditory and phonetic processes in place perception for stops. *Perception and Psychophysics, 34:* 560-568.

Schraudolph, N.N., Dayan, P. and Sejnowski, T.J. (1994). Temporal difference learning of position evaluation in the game of Go. In J.D. Cowan, G. Tesauro and J. Alspector (Eds.), *Advances in Neural Information Processing Systems, 6:* 817-824. San Mateo, CA: Morgan Kaufmann.

Schroeder, M.R., Atal, B.S., and Hall, J.L. (1979). Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In B. Lindblom and S. Ohman, *Frontiers of Speech Communication Research,* pp. 217-229. London: Academic Press.

Segui, J., Dupoux, E. and Mehler, J. (1990). The role of the syllable in speech segmentation, phoneme identification, and lexical access. In G.T.M. Altmann (Ed.), *Cognitive models of speech processing,* pp. 263-280. Hillsdale, NJ: Lawrence Erlbaum

Seneff, Stephanie (1988). A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics, 16:* 55-76.

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.

Shvachkin, N. (1948/73). The development of phonemic speech perception in children. In C.A. Ferguson and D. Slobin (Eds.), *Studies in child language development,* pp. 91-127. New York: Holt, Rinehart & Winston.

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

Singh, S.P. (1992b). The efficient learning of multiple task sequences. In J.E. Moody et. al. (eds.), *Advances in Neural Information Processing Systems, 4:* 251-258. San Mateo, CA: Morgan Kaufmann.

Singh, S.P. (1992c). Reinforcement learning with a hierarchy of abstract models. In *Proceedings of the Tenth National Conference on Artificial Intelligence,* pp. 202-207. San Mateo, CA: Morgan Kaufmann.

Singh, S.P. (1992d). Scaling reinforcement learning algorithms by learning variable temporal resolution models. In *Proceedings of the Ninth International Conference on Machine Learning,* pp. 406-415. San Mateo, CA: Morgan Kaufmann.

Slobin, D.I. (1979). Psycholinguistics (2nd Edition). Glenview, IL: Scott, Foresman and Company.

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.

Smolensky, P. (1994). Commentary on paper by L. Gleitman. Presented at the Sixteenth Annual Conference of the Cognitive Science Society, Atlanta, GA. August 15, 1994.

Spencer, A. (1986). Towards a theory of phonological development. *Lingua, 68:* 3-38.

Stampe, D. (1973). *A dissertation on natural phonology.* Ph.D. Dissertation. University of Chicago. Reprinted by New York: Garland Publishing Co.

Stevens, K.N. (1971). Airflow and turbulence noise for fricative and stop consonants: static considerations. *Journal of the Acoustical Society of America, 50:* 1180-1192.

Stevens, K.N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17:* 3-45.

Stone, M., Faber, A., Raphael, L.J. & Shawker, T.H. (1992). Cross-sectional tongue shape and linguopalatal contact patterns in [s], [S], and [l]. *Journal of Phonetics, 20:* 253-270.

Strange, W., Jenkins, J.J. and Johnson, T.L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America, 74:* 695-705.

Sussman, H.M, Minifie, F.D., Stoel-Gammon, C., Buder, E., and Smith, J. (1994). Developmental changes in C-V interdependencies: canonical babbling vs. early word attempts. Paper presented at the 19th Annual Boston University Conference on Language Development, November 6, 1994.

Sutton, R.S. (1984). *Temporal credit assignment in reinforcement learning*. Ph.D. Thesis (Technical Report COINS TR 84-02). Amherst, MA: University of Massachusetts, Computer and Information Sciences.

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

Sutton, R.S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamical programming. In *Proceedings of the Seventh International Conference on Machine Learning,* pp. 216-224. San Mateo, CA: Morgan Kaufmann.

Sutton, R.S. & Barto, A.G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks,* Chapter 12, pp. 497-538. Cambridge, MA: MIT Press.

Tallal, P., Sainburg, R.L. and Jernigan, T. (1991). The neuropathology of developmental dysphasia: Behavioral, morphological, and physiological evidence for a pervasive temporal processing disorder. *Reading and Writing, 3:* 363-377.

Tan, M. (1993) Multi-agent reinforcement learning: independent vs. cooperative agents. In *Machine Learning: Proceedings of the Tenth International Conference,* San Mateo, CA: Morgan Kaufmann.

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.

Tesauro, G.J. (1992). Practical issues in temporal difference learning. *Machine Learning, 8:* 257-279.

Tham, C.K. & Prager, R.W. (1993). Reinforcement learning methods for multi-linked manipulator obstacle avoidance and control. *IEEE Asia-Pacific Workshop on Advances in Motion Control.* Singapore, July 15-16, 1993.

Tham, C.K. & Prager, R.W. (1994). A modular Q-learning architecture for manipulator task decomposition. In *Machine Learning: Proceedings of the Tenth International Conference*.

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. (1990). 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.

Touretzky, D.S. and Wheeler, D.W. (1991). Exploiting syllable structure in a connectionist phonology model. In R.P. Lippmann, J.E. Moody, and D.S. Touretsky (Eds.), *Advances in Neural Information Processing Systems, 3:* 612-618. 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,* vol. 2, pp. 281-334. Stanford: Stanford University Press.

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. (1993). Variable paths to early word production. *Journal of Phonetics, 21:* 61-82.

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

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.

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

Wagner, R.K. and Torgesen, J.K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin, 101:* 192-212.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-37,* 328-393.

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

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

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

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

Weizenbaum, J. 1976. *Computer power and human reason: from judgement to calculation.* San Francisco: Freeman.

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

Whalen, D. H. (1990). Coarticulation is largely planned. *Journal of Phonetics, 18:* 3-35.

Whalen, D.H., Levitt, A.G. and Wang, Q. (1991). Intonational differences between the reduplicative babbling of French- and English-learning infants. *Journal of Child Language, 18:* 501-516.

Whitehead, S.D. (1991) A complexity analysis of cooperative mechanisms in reinforcement learning. In *Proceedings of AAAI-91:* 607-613.

Whitehead, S.D. (1992). *Reinforcement learning for the adaptive control of perception and action.* Ph.D. Thesis. Rochester, NY: University of Rochester, Department of Computer Science.

Wickelgren, W.A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review, 76:* 1-15.

Williams, R.J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning, 8:* 229-256.

Williams, R.J. & Peng, J. (1989). Reinforcement learning algorithms as function optimizers. *Proceedings of the 1989 International Joint Conference on Neural Networks,* II-89-95.

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