Learning to Reach Via Corrective Movements

Andrew H. Fagg[1] Andrew G. Barto[1]
James C. Houk[2]

{fagg | barto}@cs.umass.edu
[1]Department of Computer Science
University of Massachusetts
Amherst, MA 01003-4610

houk@casbah.acns.nwu.edu
[2]Department of Physiology
Northwestern University School of Medicine
Chicago, IL 60611-3008

Abstract

When infants begin to perform goal-directed reaching between approximately 4 and 5 months-of-age, the kinematics of their reaches show multiple accelerations and decelerations of the hand, which appear to reflect a correcting series of submovements. Although each submovement is an inaccurate correction, the sequence of submovements is often successful in reaching the target. By what process does the infant's reach controller come to eventually produce adult-like reaching behavior, characterized by bell-shaped velocity profiles and only occasional corrections? We suggest that motor command parameters are tuned in a supervised learning fashion using training information derived from proprioceptive feedback received during corrective movements. Although this is not high-quality training information, it is sufficient to tune motor command parameters in order to bring a dynamic arm quickly and accurately to the target, as we show in a series of simulations. We also suggest that additional parameters can be tuned to satisfy constraints other than end-point constraints through a reinforcement learning process.

1 Introduction

When infants begin to perform goal-directed reaching between approximately 4 and 5 months-of-age, the kinematics of their reaches show multiple accelerations and decelerations of the hand, which appear to reflect a correcting series of submovements. Although each submovement is an inaccurate correction, the sequence of submovements is often successful in reaching the target (von Hofsten, 1979; Berthier, 1997). By the end of the second year of life, most infants show fairly smooth and precise movements toward targets.

These and other observations made by researchers studying the development of human motor control raise many questions about the mechanisms involved in the acquisition of motor skills. Although control engineering, especially its use in robotics, provides essential help in understanding a range of possibilities, we know of no conventional control architecture that is consistent with what developmental psychologists know about human motor development. In this paper, we describe a model of the development of reaching that is informed by salient features of infant motor development, as well as by features of biological motor control in general.

Key aspects of our model are the following: 1) submovements are generated by motor commands characterized by relatively few parameters, so that no detailed trajectory planning is employed; 2) trajectory details are strongly influenced by the motor plant, which includes muscles and spinal reflexes with important nonlinear properties; 3) there is a built-in ability to make target-directed movements, but this ability is initially extremely crude; and 4) learning depends on several types of learning feedback, including proprioceptive feedback generated during corrective movements as well as evaluative feedback assessing the overall quality of movements with respect to multiple criteria. This implies that learning is a combination of supervised learning and reinforcement learning (RL). Additionally, the model has connections to a developing model of how the cerebellum and other brain structures are involved in regulating movement (e.g., Houk and Barto, 1992; Houk *et al.*, 1996; Barto *et al.*, submitted), although we do not have the space to describe these connections here. We present results of our initial efforts to assess the computational plausibility of these features.

2 A Pulse-Step Representation of the Reach Controller

When adults reach in the horizontal plane to specified targets, several regularities are observed in the trajectory of the hand. First, the hand speed exhibits a smooth, bell-shaped profile, and when accuracy requirements are low, the profile is symmetric about its peak. Second, the hand follows a nearly straight path between the initial and target positions. These regularities have been interpreted by many to mean that prior to movement execution, a detailed trajectory of the movement is planned and then transformed into a time series of muscle activation patterns that produce the desired movement (e.g., Kawato *et al.*, 1987; Kawato and Gomi, 1993; Schweighofer, 1995). Another possibility, which we explore here, is that the regularities are due in part to the dynamical properties of the limb and muscle, and that a controller need only specify a small number of control parameters. Furthermore, a controller need not explicitly represent an explicit plan of the desired behavior prior to movement execution, e.g., a trajectory in Cartesian or joint space.

One minimal representation of a two-joint reach controller consists of a pair of pulse-step joint commands (Figure 1). The pulse component of the command is responsible for initiating movement of a joint; the step specifies a region (not a unique point) in which the joint will stop moving. Pulse-step control is known to be used in the generation of rapid eye movements (Robinson, 1975) and has been suggested as a reasonable approximation for limb movement control in cat (Ghez, 1979; Ghez and Martin, 1982). In the model, the control outputs are specified by a set of seven parameters: the pulse and step heights for the shoulder and elbow (P_s, P_e, S_s, S_e), the length of the shoulder and elbow pulses (T_2 and T_3), and the relative onset time of the two pulses (T_1). These pulse-step commands are generated in a purely open-loop mode only for the sake of simplicity. In a more realistic model, the commands could be influenced by appropriately delayed feedback (Barto *et al.*, submitted).

The motor plant is a simulated, two degree-of-freedom, planar arm. The dynamic parameters were selected to be consistent with an adult arm (Gribble *et al.*, 1998). The shoulder and elbow joints are actuated by two pairs of opposing muscles. The muscle model combines a Hill-based lumped parameter model (Winters, 1990) with aspects of Feldman's λ-model (Feldman, 1966; Gribble *et al.*, 1998), and includes a nonlinear damping term mediated by the stretch reflex, as described by Gielen and Houk (1984, 1987) and Wu *et al.* (1990). The voluntary pulse-step commands set the reflex thresholds for each pair of muscles. For a given motor signal, the two opposing muscles produce a unique joint equilibrium position. However, nonlinear dynamics of the stretch reflex create a *stiction region* around the equilibrium. Once the joint enters this region, it slows rapidly, effectively halting the joint's movement. This formulation of muscle/reflex dynamics has the advantage of removing from the controller's responsibility the generation of braking bursts of muscle activity to halt movement. In effect, the stretch reflex generates braking antagonist activity. This formulation also allows rapid movements while limiting the potential for oscillations about the endpoint. More details of the muscle/reflex model may be found in Barto *et al.* (submitted) and Fagg *et al.* (1998).

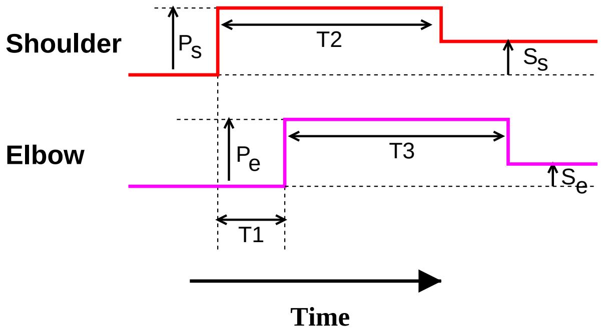


Figure 1: A Pulse-Step Representation of the Reach Control Signal. The pulse is responsible for initiating movement; the step component roughly specifies where movement should stop.

In the sections that follow, we describe two algorithms for setting the pulse-step parameters such that the target is accurately reached and a path straightness criterion is satisfied.

3 Supervised Learning

Berthier *et al.* (1993) suggested that proprioceptive feedback from individual corrective movements could provide information as to how to update previously-generated movements. Suppose that a movement is unsuccessful at reaching the target, which results in the generation of a corrective movement. The direction of the corrective movement (as measured by either the muscle activation pattern, or indirectly via proprioceptive afferents) provides training information to the controller as to how it should update its parameters for the preceding movement. Importantly, the corrective movements do not have to be accurate; they only need to move the hand, on the average, closer to the target. This can be done, as we describe below, without any consideration of the arm's dynamics.

A particular implementation of this approach is illustrated in Figure 2. The behavioral protocol is as follows: the arm is placed in an initial configuration, and a target location is presented. As a function of this information, the Supervised Learning module (*S-Learner*) selects the pulse heights (P_s and P_e). The remaining parameters are held fixed: $T_1 = 0$, $T_2 = T_3 = 150$ msec, and S_s and S_e are selected such that the joint equilibrium position is aligned with the target. Note that setting the step heights in this manner does not require knowledge of the arm dynamics, nor does this guarantee that the arm will stop at the target, due to the stiction behavior of the opposing muscles. The pulse-step commands are then executed, and the arm is allowed to come to rest. If the arm stops moving at a point far from the target, then the *Teacher* produces pulses of muscle activity, resulting in a sequence of corrective movements that bring the arm to the target. These pulses are also used to update the pulse parameters specified by the S-Learner. If the teacher produces a correction that primarily involves flexion of the elbow, for example, then the elbow pulse height (P_e) is increased in magnitude. On subsequent trials, this has the effect of producing a larger elbow flexion movement (or a smaller extension movement). Reaching trials are repeated until the S-Learner is able to reach the target without requiring correction.

The pulses generated by the Teacher are fixed duration, with magnitudes that are selected heuristically. Karst and Hasan (1991) showed that a relatively simple rule could be used to describe the initial recruitment of arm muscles as a function of the position of the target relative to the hand, without taking into account the dynamics of the arm. In the model, the relative activation level of the shoulder and elbow muscles is derived from the target's position in a visual coordinate system with origin at the wrist and aligned along the forearm (Figure 3). Target locations along the forearm axis are translated into pure shoulder movements, whereas those located along the perpendicular axis result in pure elbow movements. Off-axis targets recruit a mixture of shoulder and elbow muscles.

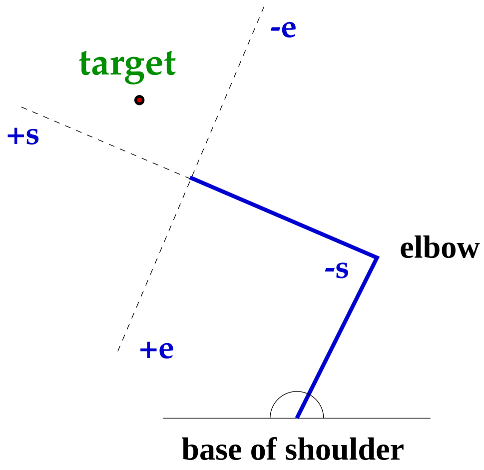


Figure 3: Heuristic Method for Generating Corrective Movements. Targets located in the half-plane denoted by $+e$ recruit elbow flexors (which drive positive accelerations of the elbow joint, i.e. $\ddot{\theta}_e > 0$); the $-e$ half-plane recruits elbow extensors; $+s$: shoulder flexors, and $-s$: shoulder extensors.

Prior to learning, the Teacher is responsible for producing all movements, as illustrated in Figure 4A. The initial position of the arm is located in the top-right corner of the workspace; the target is indicated by the "+". The path followed by the wrist is shown by the series of dots from the initial position to the target. Areas of high dot density indicate regions in which the arm slowed. Figure 4B shows the the hand speed profile (*Vel*), and the muscle activation patterns for the Shoulder and Elbow. After 40 repetitions of the same reach, the Supervised Learning controller has discovered a set of pulse parameters that brings the arm to the target without the need for corrections (Figure 5). In addition, the hand speed profile is smooth and has a single peak.

Figure 6 demonstrates the post-learning behavior of the arm for a variety of initial and target positions. For each initial/target pair, the pulse parameters were learned independently (no mechanism has been implemented for generalization across similar reaches, although this is a straightforward extension of the model). Note that many of the paths con-

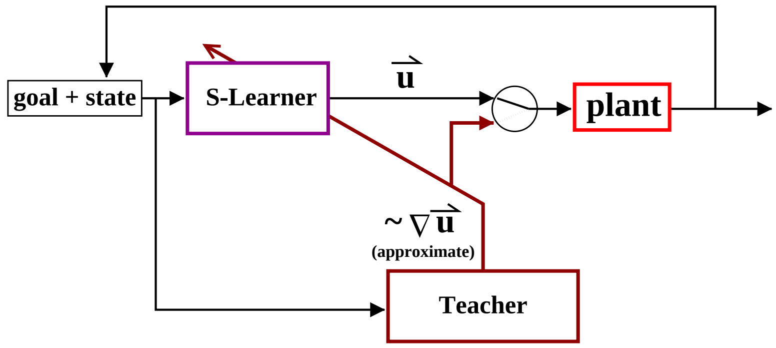
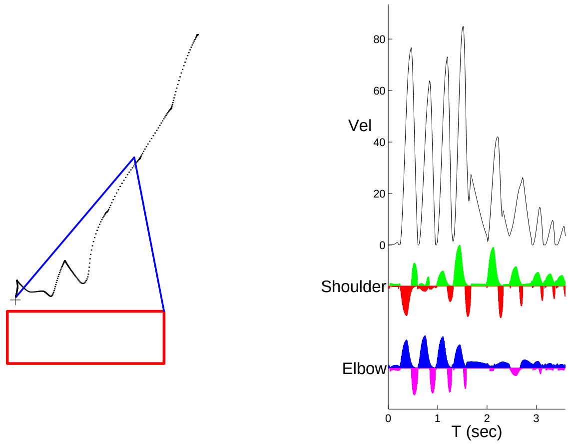
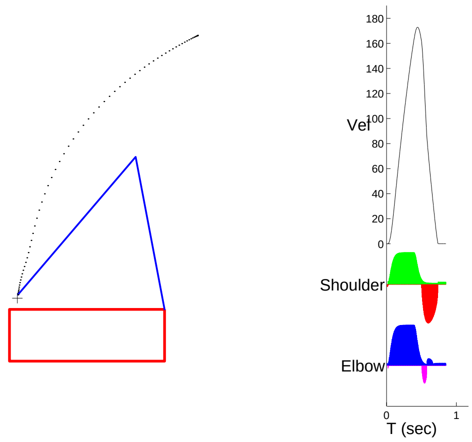


Figure 2: A Supervised Learning Architecture. The plant is driven by control signals produced by either the Supervised Learning (*S-Learner*) module, or the *Teacher* (only one module produces control signals at any one time). The S-Learner first makes an attempt at reaching the target. If unsuccessful, the Teacher produces a sequence of corrective signals that are used to both drive the arm closer to the target on the current trial, and to update the pulse-step control signals represented by the S-Learner.



A B

Figure 4: Behavior of the Arm Prior to Learning. The hand speed profile (Vel) is measured in cm/sec. Shoulder and elbow muscle activation patterns (representations of EMGs) are read as follows: excursions upward from baseline represent flexor activation; excursions downward represent extensor activation. This measure of muscle activity includes both centrally-generated and stretch reflex components.



A B

Figure 5: Behavior of the Arm After Learning. The movement is generated entirely by the *S-Learner* pulse-step controller. The resulting hand speed profile is smooth and has a single peak.

tain "hooks" near the end of movement. In each of these cases, one joint (typically the shoulder) has completed its motion long before the other. In addition, the hooked paths typically exhibit two peaks in the hand speed profile.

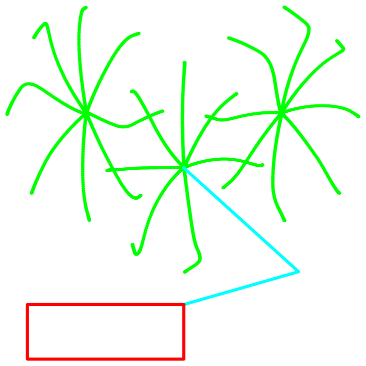


Figure 6: Post-Learning Paths for 3 Initial Positions (Centers) and 23 Targets. Note the significant hooks that occur at the end of many of the movements.

4 A Hybrid Supervised and Reinforcement Learning Approach

The undesirable aspects of the trajectories described above can be addressed by adjusting additional pulse-step parameters (specifically, T_1 and T_3 of Figure 1). The critical question is how to best select these parameters given the observed behavior of the arm on one or more trials. It is possible to address the "hooking problem" by extending the supervised learning method in the following manner. If the arm reaches the target *without correction*, but the shoulder completes its motion before (after) the elbow, then T_3 is increased (decreased) by a fixed duration. If T_3 is updated, then another trial is executed, and the pulse-step parameters (P_s, P_e, and T_3) are adjusted as necessary.

Although this heuristic method can minimize the observed hooking behavior, the resulting paths are approximately straight in joint space, which implies that the paths are often significantly curved in Cartesian space. Hollerbach and Atkeson (1987) suggested that approximately straight Cartesian movements can be achieved by staggering the onset of joint movement, which can be accomplished in this model by adjusting the T_1 parameter. However, it is not apparent that there is a simple rule for extracting a direction of adjustment of the T_1 parameter from the observed behavior of the arm on a single trial (specifically, the straightness of the path). While we could in principle compute the appropriate Jacobian from a model of the arm (cf. Jordan *et al.*, 1994), it is much simpler to use an RL-type search process to discover the T_1 that results in the straightest hand path.

One possible architecture is shown in Figure 7. Two separate learning modules, the *S-Learner* and the *R-Learner*, specify orthogonal components of the pulse-step controller (specifically, $u^S = \{P_s, P_e, T_3\}$ and $u^R = \{T_1\}$). The S-Learning module utilizes the method described above to discover the appropriate u^S for a given situation, while in parallel, the R-Learning module makes use of an RL approach (Sutton and Barto, 1998) to search for the optimal u^R.

The learning protocol is as follows: For a given initial position and target, the R-Learning module selects a particular u^R. As a function of the initial and target positions (and possibly u^R), the S-Learning module selects a u^S, and the motor signals are generated. If the target is not reached or the two joints do not complete their motions simultaneously, the Teacher provides a heuristic update to u^S, and produces a sequence of corrective movements that bring the arm to the target. Reaching trials are repeated, with the same u^R, until changes become sufficiently small. The resulting trajectory is evaluated by the Critic as a function of its deviation from the straight-line path between the initial and target positions (the reward, R, is highest for no deviation from the straight path). Based upon this evaluation, the R-Learner selects a new u^R, and the supervised learning process is repeated. An important distinction between these two learning systems is that the Teacher informs the S-Learner as to *how* to update its motor signal, whereas the Critic only provides an evaluation of the resulting behavior and not a directional change to the R-Learner's motor signal.

For the experiments described below, the R-Learner makes use of a hierarchical search in order to discover the appropriate u^R. At the first level of the search, a coarse sampling of T_1's are evaluated. The best performing T_1 (maximum R) becomes the center of the search at the next level, around which a finer sampling of T_1's are selected. This process is

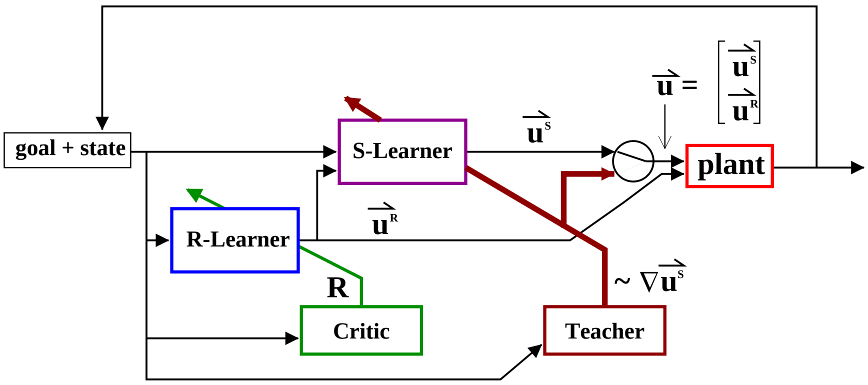


Figure 7: A Hybrid Supervised-Reinforcement Learning Architecture. The two learning controllers (*S-Learner* and *R-Learner*) specify separate components of the motor signals. As in Figure 2, the Teacher produces corrective movements, which are also used to tune the motor signals produced by the S-Learner (u^S). The *Critic* provides an evaluation (R) of a reach based on the straightness of the path.

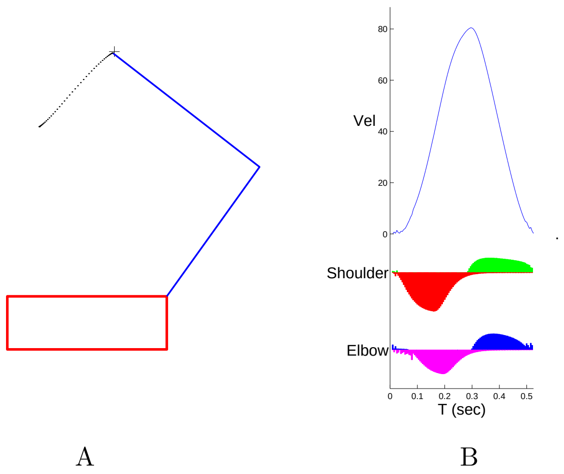


Figure 8: Behavior of the Arm After Hybrid Learning. Note that the onset of the elbow pulse is delayed relative to the shoulder, producing a relatively straight path in Cartesian space.

repeated until the finest granularity is reached.

In practice, between 200 and 700 movement trials are required to find the appropriate set of parameters for a single starting point/target pair. The resulting paths are often nearly straight, and the velocity profiles are smooth and exhibit single peaks. (Figures 8 and 9).

5 Discussion

Adult reaching movements exhibit regularities in the form of straight hand paths and smooth, bell-shaped hand speed profiles. To produce these regularities, in what form does the reach controller specify motor signals, and what training information is necessary to tune these signals with experience? In our model, the controller represents the motor signals by specifying the height and timing parameters of a pulse-step waveform for each joint, making no use of an explicit plan for the desired trajectory to be followed.

Despite the low dimensionality of the control signal representation, it is possible to produce regularities similar to those of adult reaches, due in part to the dynamical behavior of the muscles and the arm. This approach is similar to that of Karniel and Inbar (1997), a key difference being that only the agonist muscle bursts are specified directly in our model. The braking antagonist burst arises naturally from the stretch reflex built into our model. Their learning procedure also relies on a systematic variation of all of the control parameters to estimate an error gradient, in order to update control parameters.

Our model, on the other hand, assumes the existence of a crude mechanism for the generation of corrective movements. While it is easy to *initiate* movement in a generally correct direction, it is much harder to accurately reach a target endpoint with a fast and smooth movement. Our model shows that

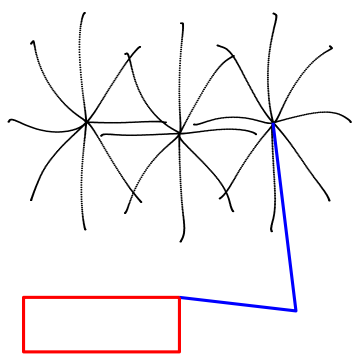


Figure 9: Post-Learning Paths (Hybrid Case) for a Variety of Initial Positions and Targets. By making use of the T_1 and T_3 parameters, it is possible to achieve nearly straight paths (compare with Figure 6).

crude corrective movements not only allow successful reaching on every trial (although early reaches may require many corrections), but they can also provide training information useful for updating a subset of controller parameters so that faster and smoother movements result. This process produces a progression from multi-segmented reaches qualitatively similar to those of infants to more skillful reaches similar to those of adults. An additional RL mechanism adjusts pulse timing parameters so that the hand follows a relatively straight path. It would have been possible to use RL to adjust all of the controller's parameters, but this would have required a more difficult search in a higher-dimensional space. We think it is plausible that infants use RL to tune aspects their movements, although it is unlikely that they use a straightness criterion as we have done here. It is more likely that an infant's reinforcement function depends on factors such as movement time and accuracy.

Our learning approach is related to Kawato's *feedback error learning* in that learning is guided by movement produced by a feedback controller external to the learning module (Kawato *et al.*, 1987; Kawato and Gomi, 1993; Schweighofer, 1995). Our model differs from this work in two respects. First, feedback in our model is not provided continuously, rather it is only available sporadically. Second, our model does not require the use of a high-quality reference trajectory from which the corrective movements are computed. Although Cartesian path information is used by the RL component in adjusting the pulse timing parameters, this information is utilized *after* the reaching movement, and not during its planning or execution.

Although the controller in the present model is purely open-loop and learns to perform a movement from a single position to a single target, the learning approach is more generally applicable at two levels. First, it is possible to learn the pulse-step parameters as a function of the current state of the arm and target position, allowing the model to learn reaching movements over the entire workspace. Second, in the context of a cerebellar learning model, it is possible to learn to combine contextual information, motor efference copy, and delayed sensory inputs in order to compute the outgoing motor commands in an on-line fashion (Fagg *et al.*, 1997; Barto *et al.*, submitted).

Acknowledgments

This work is supported by the National Institute of Mental Health (grant NIH 1-50 MH 48185-02) and by the National Science Foundation (grant NSF IRI-9720345).

References

Barto, A. G., Fagg, A. H., Sitkoff, N., and Houk, J. C. (1998). A cerebellar model of timing and prediction in the control of reaching. *Submitted to Neural Computation.*

Berthier, N. E. (1997). Analysis of reaching for stationary and moving objects in the human infant. In Donohoe, J., editor, *Neural Network Models of Cognition-Biobehavioral Foundations*, pages 283–301.

Berthier, N. E., Singh, S. P., Barto, A. G., and Houk, J. C. (1993). Distributed representations of limb motor programs in arrays of adjustable pattern generators. *Journal of Cognitive Neuroscience*, 5:56–78.

Fagg, A. H., Sitkoff, N., Barto, A. G., and Houk, J. C. (1997). A model of cerebellar learning for control of arm movements using muscle synergies. In *Proceedings of the IEEE Symposium on Computational Intelligence in Robotics and Automation*, pages 6–12.

Fagg, A. H., Zelevinsky, L., Barto, A. G., and Houk, J. C. (1998). A pulse-step model of control for arm reaching movements. In *Proceedings of the 1998 Meeting of the Society for the Neural Control of Movement*.

Feldman, A. (1966). Functional tuning of the nervous system with control of movement or maintenance of a steady posture. II Controllable parameters of the muscle. *Biophysics*, 11:565–578.

Ghez, C. (1979). Contributions of central programs to rapid limb movement in the cat. In Asanuma, H. and Wilson, V. J., editors, *Integration in the Nervous System*, pages 305–320. Igaku-Shoin, Tokyo.

Ghez, C. and Martin, J. H. (1982). The control of rapid limb movment in the cat. III. Agonist-antagonist coupling. *Experimental Brain Research*, 45:115–125.

Gielen, C. C. A. M. and Houk, J. C. (1984). Nonlinear viscosity of human wrist. *Journal of Neurophysiology*, 52:553–569.

Gielen, C. C. A. M. and Houk, J. C. (1987). A model of the motor servo: Incorporating nonlinear spindle receptor and muscle mechanical properties. *Biological Cybernetics*, 57:217–231.

Gribble, P. L., Ostry, D. J., Sanguineti, V., and Laboissière, R. (1998). Are arm control signals required for human arm movement? *Journal of Neurophysiology*, 79(3):1409–1424.

Hollerbach, J. M. and Atkeson, C. G. (1987). Deducing planning variables from experimental arm trajectories: Pitfalls and possibilities. *Biological Cybernetics*, 56:67–77.

Houk, J. C. and Barto, A. G. (1992). Distributed sensorimotor learning. In Stelmach, G. E. and Requin, J., editors, *Tutorials in Motor Behavior II*, pages 71–100. Elsevier Science Publishers B. V., Amsterdam, The Netherlands.

Houk, J. C., Buckingham, J. T., and Barto, A. G. (1996). Models of the cerebellum and motor learning. *Behavioral and Brain Sciences*, 19:368–383.

Jordan, M. I., Flash, T., and Arnon, Y. (1994). A model of the learning of arm trajectories from spatial deviations. *Journal of Cognitive Neuroscience*, 6:359–376.

Karniel, A. and Inbar, G. F. (1997). A model of learning human reaching movements. *Biological Cybernetics*, 77:173–183.

Karst, G. M. and Hasan, Z. (1991). Initiation rules for planar, two-joint arm movements: Agonist selection for movements throughout the work space. *Journal of Neurophysiology*, 66(5):1579–1593.

Kawato, M., Furukawa, K., and Suzuki, R. (1987). A hierarchical neural-network model for control and learning of voluntary movement. *Biological Cybernetics*, 57:169–185.

Kawato, M. and Gomi, H. (1993). Feedback–error–learning model of cerebellar motor control. In Mano, N., Hamada, I., and DeLong, M. R., editors, *Role of the Cerebellum and Basal Ganglia in Voluntary Movement*, pages 51–61. Elsevier Science Pub. B. V.

Robinson, D. A. (1975). Oculomotor control signals. In Lennerstrand, G. and Bach-y-rita, P., editors, *Basic Mechanisms of Ocular Motility and their Clinical Implications*, pages 337–374. Pergamon Press, Oxford.

Schweighofer, N. (1995). *Computational Models of the Cerebellum in the Adaptive Control of Movements*. PhD thesis, University of Southern California, Los Angeles, CA.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

von Hofsten, C. (1979). Development of visually directed reaching: The approach phase. *Journal of Human Movement Studies*, 5:160–168.

Winters, J. M. (1990). Hill-based muscle models: A systems engineering perspective. In Winters, J. M. and Woo, S. L.-Y., editors, *Multiple Muscle Systems: Biomechanics and Movement Organization*, pages 69–93. Springer-Verlag, New York.

Wu, C. H., Houk, J. C., Young, K. Y., and Miller, L. E. (1990). Nonlinear damping of limb motion. In Winters, J. M. and Woo, S. L.-Y., editors, *Multiple Muscle Systems: Biomechanics and Movement Organization*, pages 214–235. Springer–Verlag, New York.