Reinforcement Learning of Adaptive Longitudinal Control for Dynamic Collaborative Driving

Luke Ng

Department of Mechanical and Mechatronics Engineering,
University of Waterloo,
200 University Ave W., Waterloo, Ontario, N2L 3G1, Canada
Fax: (519) 888-4333
Email: l4ng@engmail.uwaterloo.ca

Christopher M. Clark

Computer Science Department,
California Polytechnic State University
San Luis Obispo, CA, 93407, USA
Fax: (805) 756-2956
Email: cmclark@calpoly.edu

Jan P. Huissoon

Department of Mechanical and Mechatronics Engineering,
University of Waterloo,
200 University Ave W., Waterloo, Ontario, N2L 3G1, Canada
Fax: (519) 888-4333
Email: jph@uwaterloo.ca

Abstract: Dynamic collaborative driving involves the motion coordination of multiple vehicles using shared information from vehicles instrumented to perceive their surroundings in order to improve road usage and safety. A basic requirement of any vehicle participating in dynamic collaborative driving is longitudinal control. Without this capability, higher-level coordination is not possible. Each vehicle involved is a composite nonlinear system powered by an internal combustion engine, equipped with automatic transmission, rolling on rubber tires with a hydraulic braking system. This paper focuses on the problem of longitudinal motion control. A longitudinal vehicle model is introduced which serves as the control system design platform. A longitudinal adaptive control system which uses Monte Carlo Reinforcement Learning is introduced. The results of the reinforcement learning phase and the performance of the adaptive control system for a single automobile as well as the performance in a multi-vehicle platoon is presented.

Keywords: autonomous robotics; mobile robots; motion control; adaptive cruise control; collaborative driving; vehicle dynamics; vehicle simulation; artificial intelligence; machine learning; reinforcement learning; adaptive control.

Reference to this paper should be made as follows: Ng, L, Clark, C.M., Huissoon, J.P. (2008) 'Reinforcement learning of adaptive longitudinal control for dynamic collaborative driving', Int. J. Vehicle Information and Communication Systems, Vol. X, No. Y, pp.000–000.

L. Ng, C.M. Clark, J. P. Huissoon

Biographical notes: Luke Ng is a doctoral candidate at the Lab for Autonomous and Intelligent Robotics, Dept. of Mechanical and Mechatronics Engineering, University of Waterloo.

Chris M. Clark is an Assistant Professor at the Computer Science Dept., California Polytechnic State University, San Luis Obispo, CA.

Jan P. Huissoon is a Professor at the Dept. of Mechanical and Mechatronics Engineering, University of Waterloo. He is the currently the Department Deputy Chair and is a Professional Engineer.

1. Introduction

In major cities throughout the world, urban expansion is leading to an increase of vehicle traffic flow. The adverse effects of increased vehicle traffic flow include traffic congestion, driving stress, increased vehicle collisions, pollution, and logistical delays. Once traffic flow surpasses the capacity of the road system, it ceases to become a viable transportation option. One solution is to build more roads; another is to build better vehicles– vehicles that can negotiate traffic, coordinate with other similar 'thinking' vehicles to optimize their speeds so as to arrive at their destination safely and efficiently. This is the concept behind Dynamic Collaborative Driving, an automated driving approach where multiple vehicles dynamically form groups and networks, sharing information in order to build a dynamic representation of the road to coordinate travel.

Ultimately our research goal is to create a decentralized control system capable of performing dynamic collaborative driving which is scalable to a large number of vehicles, can be implemented on any vehicle and in any environment. However, before we can deal with the issue of coordination, basic control of the vehicle must be achieved. Therefore, the focus of this paper is the basic problem of longitudinal motion control, sometimes referred to as adaptive cruise control.

Research in automated driving in the United States during the 1990s was conducted under the PATH project (Partners for Advanced Transit and Highways). PATH introduced the concept of platooning (Varaiya 1993; Shladover et al 1993; Hedrick et al 1994), where vehicles in groups of 10-25 cars travel in tight vehicle-string formations. The most basic level of control in platooning is longitudinal control, also referred to as autonomous intelligent cruise control (AICC). Ioannou and Chien (1993) describe an AICC system for automatic vehicle following, which is a stand-alone longitudinal control system using a linear vehicle following model. Raza and Ioannou (1997) implemented the AICC system on a real vehicle and evaluated it during Demo '97 to verify the performance obtained under simulation.

Studies in the mid 1990s at UC Berkeley (Maciuca and Hedrick 1995; Swaroop & Hedrick 1994) focused on using sliding mode control to address the nonlinearities of longitudinal vehicle dynamics. The studies addressed both vehicle dynamics simulation, string stability of linear formations and nonlinear control. Rajamani et al (2000) implemented sliding surface control to longitudinal control during Demo '97. At Demo 2000, Kato et al (2002) demonstrated an adaptive proportional control law for longitudinal control. Recently, Zhang and Ioannou (2005) proposed an adaptive control approach to vehicle following with variable time headways, using a simplified first order

linear vehicle model. The control system guarantees closed-loop system stability, and regulates the speed and separation errors towards zero when the lead vehicle is at a constant speed. Khatir and Davision (2006) revisited linear control approaches by proposing a linear PID controller for longitudinal and lateral control assuming a simplified 6^{th} order linear model, where the vehicle is modelled as a bicycle to further simplify vehicle dynamics.

Due to the high costs associated with procuring large numbers of vehicles and the safety issues involved, full-scale vehicle studies can only be conducted through large scale research projects in association with governments and automobile manufacturers such as Demo '97 (Thorpe et al 1997; Tan et al 1998; Rajamani et al 2000) and in Japan during Demo 2000 (Tsugawa et al 2000; Kato et al 2002). In Canada, smaller projects have used mobile robots to model cars (Michaud et al 2006), however the cost and complexity associated with these mobile robot studies can also be quite high. In addition the vehicle dynamics of a mobile robot platform are significantly different from those of full-sized automobiles thereby limiting the applicability of those results.

Alternatively, simulation studies can be developed faster, they are more flexible, cost effective, have better repeatability and explore situations not easily achieved in reality. In 1989, the National Highway Traffic Safety Administration (NHTSA) began researching the use and construction of a new state-of-the-art driving simulator, the National Advanced Driving Simulator (NADS) (Haug 1990). Since then, NADS has been used as a substitute for actual vehicle testing. The NHTSA's Vehicle Research and Test Center (VRTC) provides vehicle data for several vehicles such as the 1997 Jeep Cherokee (Salaani and Heydinger 2000), which can be used to validate simulations. With the adoption of high fidelity simulation on modern computers, simulation has become the dominant method for study in this field.

Therefore, our methodology involves first creating an accurate vehicle model to be used both in the process of design and validation of the control system. The use of a computer model can be considered a Computer Aided Engineering (CAE) approach to control design allowing the designer to assess the performance of the control system and predict its limits. We proceed with a description of the vehicle dynamics model followed by an explanation of the longitudinal control system's design, then, the results of the learning process and the evaluation of the system's performance are presented.

2. Vehicle Dynamics Modelling

The basis of our simulation has its roots in the late 1980's to the late 1990's. A significant amount of research was conducted at the Vehicle Dynamics Laboratory at the University of California at Berkeley by Hedrick under the PATH project. His group developed a complex numerical automobile model used to design and evaluate the performance of various controllers under certain conditions (McMahon and Hedrick 1989; Peng and Tomizuka 1991; Pham et al 1994). Later work by Pham and Hedrick (Pham et al 1997) used this model to evaluate the performance of an optimal controller for combined lateral and longitudinal control.

The vehicle model adopts many of the models used by Hedrick's group for key subsystems such as the engine, transmission, suspension and tires. However, in order to

have a simulation which can be subjected to reinforcement learning, these separate models have been integrated and modified to provide system performance throughout the entire operating range. For example, an automatic transmission system is added to allow gear shifting so that the entire speed range can be experienced. Figure 1 illustrates how each subsystem model is interconnected into a coherent model of an automobile. The following is a partial description of each of the major subsystem models and shows where the nonlinearities of the overall vehicle model originate.

2.1. Engine Model

McMahon and Hedrick (1989) describe in detail a mathematical model of a 3.4*L Ford V6* internal combustion engine. The control input to the engine is the throttle angle α, which is supplied by the throttle actuator model. The throttle actuator model is simply a first order system with a time constant of 0.050 *ms*. The output of the engine model is the engine's crankshaft speed. In addition, a feedback term from the transmission model is required in the form of the torque of the transmission pump which is connected to the engine's crankshaft.

The engine model is made up of several differential equations for each part of the combustion process. The differential equation for the gas mixture in the intake manifold is given by

$$\dot{P}_m = \left(\frac{\dot{m}}{m} \right) - 0.08873\omega \quad P_m + \left(\frac{RT_m}{V_m} \right) \quad \dot{m}_{ai} + \dot{m}_{egri} \quad KPa/s \qquad (1)$$

where P_m, and T_m are the manifold pressure and temperature. The manifold volume V_m is considered fixed at 3.4*L* or 0.0034 m^3. The mass rate of air entering the intake manifold is given by the relationship

$$\dot{m}_{ai} = MAX \cdot TC \cdot PRI \qquad kg/s, \qquad (2)$$

where $MAX = 0.335$ *kg/s* is the engine specific maximum flow rate. *TC* is the normalized throttle characteristic and is a function of throttle angle α, and *PRI* is the normalized pressure influence ratio and is a 5th order polynomial function of the pressure ratio $PR = P_m/P_{atm}$.

The differential equation for the gas mixture in the exhaust manifold is expressed as

$$\dot{m}_{ao} = \dot{m}_{ai} + \dot{m}_{egri} - \dot{m}_{egro} + \left(\frac{P_m V_m}{RT_m} \right) \left(\frac{\dot{T}_m}{T_m} - \frac{\dot{P}_m}{P_m} \right) \qquad kg/s \qquad (4)$$

Where the exhaust gas recirculation out of the exhaust manifold is described by the second order differential equation

$$\ddot{m}_{ergo} = 1.5 \times 0^{-} \quad \eta \quad \omega \quad \dot{m}_{ergi} - \dot{m}_{ergo} \qquad kg/s \qquad (5)$$

and volumetric efficiency term η_{vol} of the engine is expressed as a surface (Figure 2) with a dependence on the mass of air flow rate into the intake manifold and the rotational speed of the engine ω_e. The mass flow rate of the exhaust gas into the exhaust manifold $\dot{m}_{egri} = EGRI(P_m / P_e)$ *kg/s* is provided by a lookup which is dependent on the ratio of manifold to exhaust pressure P_m/P_e. The pressure of the air in the exhaust manifold can be determined by the relationship

$$P_e = 02 + P_m \cdot 7 \times 0^{-} \cdot \omega \quad t - \quad _{it} - .12 \qquad KPa \qquad (6)$$

where the engine speed is a function of time $\omega_e(t)$ where $t = t - \Delta_{it}$, the delayed time of the intake to torque production delay resulting from the cyclical nature of the engine. The

time delay is $\Delta_{it} = 5.48/\omega_e$ *sec*. The indicated engine torque produced is modelled in the continuous time domain for simplicity as

$$T_i = \frac{c_t \dot{m}_{ao}(t-\lambda) \cdot AFI(t-\lambda) \cdot SI(SA(t-\lambda))}{\omega_e(t-\lambda)} \quad N \cdot m \qquad (7)$$

The constant $c_t = 1175584\ N \cdot m \cdot s/kg$ is the maximum torque capability of the engine, the function $AFI(t)$, is the air/fuel influence function. The function $SI(SA) = 1.0\text{-}3.8\text{x}10^{-4}(SA(t))^2$, is the normalized spark influence function which is also a function of spark advance $SA(t)$ from *MBT* or *minimum spark advance for best torque*.

The engine speed can be determined using the following torque balance or differential equation

$$J_e \dot{\omega}_e = T_i - T_f - T_p \quad N \cdot m \qquad (8)$$

where $J_e = 0.2630\ kg \cdot m^2$ is the effective inertia of the engine and torque converter, T_i is the indicated torque produced by the engine, $T_f = T_{fric}(\dot{m}_{ao}, \omega_e)$ is the engine friction torque function and T_p is the torque converter pump torque which is modelled in the transmission subsystem model.

2.2. Transmission Model

McMahon and Hedrick (1989) describe the model of an automatic transmission subsystem (Figure 3). The transmission system connects the engine to the driveshaft where the motion from the engine is transmitted through a gear-train to the driveshaft. The engine is connected directly to the pump of the torque converter. The rotational motion of the fluid transmits the torque from the pump to the turbine. The turbine's output shaft is connected to the gear-train which is connected to the driveshaft. Since the gear-train is only connected to the engine through the transmission fluid, it is possible to change gears without disrupting the motion of the engine. The control of the gear selection of the gear train is managed by the valve body which senses hydraulic pressure and actuates servo pistons to select the proper gear ratios to optimize engine performance. The behaviour of the valve body can be modelled as discrete logic schedule which is dependent on both vehicle speed and throttle position.

There are two phases of operation for the torque converter, the high torque phase (10 and 11) experienced when changing gears and the fluid coupling phase (13). The torque equations depend on the speed ratio of the turbine and pump ω_t/ω_p, the high torque phase satisfies the relationship $\omega_t/\omega_p < 0.9$. The torque of the pump T_p and the turbine T_t are expressed using the following equations

$$T_p = 1.4325 \times 10^{-6}\omega_{e,eff}^2 + 1.21 \times 10^{-6}\omega_{e,eff}\omega_{t,eff} \qquad (9)$$

$$T_t = -1.7656 \times 10^{-6}\omega_{e,eff}^2 + 1.3107 \times 10^{-6}\omega_{e,eff}\omega_{t,eff} - 1.4323 \times 10^{-6}\omega_{t,eff}^2 \qquad (10)$$

where $\omega_{p,eff}$ and $\omega_{t,eff}$ satisfy the first order lag expressions

$$\omega_{e,eff}\tau_e + \omega_{e,eff} = \omega_e \qquad \omega_{t,eff}\tau_t + \omega_{t,eff} = \omega_t \quad rad/s \qquad (11)$$

The fluid coupling phase exists when $\omega_t/\omega_p \geq 0.9$, therefore the torques for the turbine and pump are expressed as

$$T_t = T_p = -1.7644 \times 10^{-6}\omega_e^2 + 2.0084 \times 10^{-6}\omega_e\omega_t - 5.2441 \times 10^{-6}\omega_t^2 \qquad (12)$$

Since the engine is connected directly to the pump, $\omega_e = \omega_p$, thus, the angular speed of the turbine ω_t, can be determined with the first order differential torque balance equation,

$$J_{tg}\dot{\omega}_t = T_t - K_g R_d T_s \quad N \cdot m \qquad (13)$$

where J_{tg}= 0.07 $kg{\cdot}m^2$ is the rotational inertia, R_g is the gear ratio depending on which gear is used (i.e. 0.4167, 0.6817, 1, 1.4993) and R_d = 1 is the drive gear ratio. The shaft torque T_s can be determined from the following first order differential equation

$$\dot{T}_s = K_s(R_g R_d \omega - \omega_{wf}) \qquad N{\cdot}m/s \quad (14)$$

where K_s = 6742 $N{\cdot}m/rad$ is the shaft stiffness and ω_{wf} is the angular speed of the front wheel.

2.3. Braking System Model

McMahon and Hedrick (1989) describe a simplified model to determine the braking forces to apply to each wheel. Although the brake torques are largely dependent on the hydraulic system that makes up the braking system of the vehicle (Maciuca and Hedrick 1995), a first order lag expression provides a sufficient simplified approximation to the system. A normalized brake command cmd_{brake} in the interval [0, 1] is assumed to be provided by the control system and is passed through the brake actuator model. This is simply a first order system with a time constant of 0.075 *ms*. The first order lag function, lag_{brake} which approximates the braking system is modelled with a time constant of τ = 0.072 *s*. The equation for the braking torques for the front T_{bf} and rear T_{br} are

$$T_{bf} = lag_{brake}(actuator_{brake}(s_{brake})) \cdot h_f F_{f\max} \qquad T_{br} = lag_{brake}(actuator_{brake}(s_{brake})) \cdot h_r F_{r\max} \quad (15)$$

where h_f = 0.310 *m* and h_r = 0.315 *m* are the heights to the from the ground to the front and rear axles respectively. The maximum brake force $F_{f\,max}$ and $F_{r\,max}$ occurs during wheel lock (slip λ = 1) and can be determined using the following equations

$$F_{f\max} = \mu mg \cdot (l_r + f_{roll}(\mu - f_{roll})) \qquad F_{r\max} = \mu mg \cdot (l_f + f_{roll}(\mu - f_{roll})) \quad (16)$$

where μ is the coefficient of friction between the road and the tire as specified in the tire subsystem model, m = 1573 *kg* is the mass of the automobile, g = 9.807 m/s^2 is gravity, l_f = 1.034 m and l_r = 1.491 *m* is the longitudinal distance from the center of gravity to the front and rear axles respectively and f_{roll} = 0.004908 is the coefficient of rolling resistance for the left and right tires combined.

2.4. Drive-train Model

Pham et al (1997) describe the model of the drive-train subsystem for a front wheel drive automobile. A torque balance about each wheel yields the first order differential equations for the front and the rear wheels are

$$J_{w_i}\dot{\omega}_{w_i} = \frac{1}{2}T_s - \frac{3}{10}T_b - r_{w_i}F_{x_i} \quad i = 1, 2 \qquad J_{w_i}\dot{\omega}_{w_i} = \frac{2}{10}T_b - r_{w_i}F_{x_i} \quad i = 3, 4 \quad (17)$$

where T_s is the shaft torque calculated previously in the transmission subsystem model and T_b is the total braking torque available and F_x is the longitudinal force of each tire which is calculated by the tire model. An even distribution of the shaft torque is assumed by splitting half of the shaft torque to each of the front wheels. The total available brake torque is assumed to be distributed 60% to the front and 40% to the rear wheels.

2.5. Suspension Model

Pham et al (1997) describe a simple one-dimensional quarter car model of an automotive suspension system with shock absorber and hardening spring (Peng 1992). Neglecting the small coupling terms, the suspension forces can be completely determined by the

local motion of each wheel (Tseng 1991). Let e_i be the deflection at the i^{th} suspension joint.

$$e_1 = z_0 - z + h_5\theta + l_f\theta + \frac{s_f}{2}\phi \qquad e_2 = z_0 - z + h_5\theta + l_f\theta - \frac{s_f}{2}\phi$$

$$e_3 = z_0 - z + h_5\theta - l_r\theta - \frac{s_r}{2}\phi \qquad e_4 = z_0 - z + h_5\theta - l_r\theta + \frac{s_r}{2}\phi \quad (18)$$

where z_0 is the nominal height, z is the current height of the vehicle, θ is the pitch angle, ϕ is the roll angle, $h_5 = 0.1$ *m* is the longitudinal distance from the center of gravity to pitch center, $l_f = 1.034$ *m* and $l_r = 1.491$ *m* is the longitudinal distance from the center of gravity to the front and rear axles respectively, and $s_f = 1.450$ *m* and $s_r = 1.450$ *m* are the front and rear axles respectively.

The spring force F_s and damping force F_d on each wheel is calculated using the following equations

$$F_{s_i} = C_1(e_1 + C_2 e_i^3) \quad N \qquad F_{d_i} = D_1\dot{e}_i \quad N \quad (19)$$

where $C_1 = 40000$ *N/m* and $C_2 = 40000$ m^{-4} are coefficients of the third order polynomial fit for the suspension spring and $D_1 = 10000$ *N·s/m* is the damping constant. A vertical force balance is used to determine the normal force F_N exerted on each wheel,

$$F_{N_i} = \frac{1}{2} m g \frac{l_f}{l_f + l_r} - F_{s_i} - F_{d_i} \quad N \quad (20)$$

where $m = 1573$ *kg* is the mass of the automobile and $g = 9.807$ *m/s*^{2} is gravity.

2.6. Tire Model

Pham et al (1997) describe a simplified tire model referred to as the Bakker-Pacejka model adopted from the work of Peng (1992). This model calculates the traction force resulting from the road-tire interaction based on empirical curve-fitting with experimental data for a *Yokohama P205/60R1487H* tire (Peng 1992) (see Figure 4). In this model, tire pressure, tire camber angle, and the road and tire physical parameters are fixed, but the forces generated at the tire are the functions of slip ratio λ and the tire normal force F_N.

The calculation of the slip ratio λ which is computed either for traction or braking using the following equations

$$\text{Traction: } \lambda = \frac{r_w\omega - V_x}{r_w\omega} \geq 0 \quad \text{Braking: } \lambda = \frac{r_w\omega - V_x}{V_x} < 0 \quad (21)$$

where ω_w is the rotational speed of each wheel determined in the drive-train subsystem model and the radius of the tire is $r_w = 0.304$ *m*.

According to Bakker et al (1987), road-tire interaction under non-ideal conditions can be extrapolated from the ideal curve by multiplying the ideal tire forces by the coefficient of friction μ. Typically for average freeway operation, $\mu = 0.8$, for wet road conditions $\mu = 0.6$, and for icy road conditions $\mu = 0.2$.

2.7. Vehicle Model Response

To demonstrate the performance and the validity of the vehicle dynamics simulation, the model is subjected to input signals for either the throttle or brake command and the velocity response is charted. Although the simulation is a composite of various subsystems that does not correlate to a standard vehicle (i.e. Ford V6 3.0L engine,

Yokohama 15" radial tires, Toyota Camry chassis dimensions), comparison with vehicle response data supplied by the NHTSA's Vehicle Research and Test Center (VRTC) for a 1997 Jeep Cherokee (Salaani and Heydinger 2000) is presented to illustrate analogous behaviour between an actual vehicle and simulation. It is not precision that is being compared rather accuracy in terms of behaviour.

Figure 5 shows the actual vehicle's velocity response to a throttle step input in 1st gear and the brake step response. In the throttle step, there is a smooth increase in acceleration which saturates at the top speed for the specific gear. In the brake step, from 0-6 *s* the throttle is released, this is known as power-off and results in a linear decrease in speed due to engine braking, at 6 *s* the brake is pressed and a linear decrease with a much steeper slope is seen.

A commercial mechanical simulation called Adams Car (MSC Software Corp.) is used as an intermediate validation tool by providing data for throttle and brake inputs not provided by Salaani and Heydinger (2000). The vehicle modelled in Adams Car is a high performance sports car, therefore a comparison with our *Simulation* will show the same behaviour but the responses will be slower. Both simulations are subject to the same input signals and the results are presented in Figures 6 to 8. Figures 6 shows the simulation velocity response to a throttle step input. Notice that the vehicle speed range is much larger since the simulation incorporates an automatic transmission system. The effects of the automatic transmission shifting can be seen as slight discontinuities in the response. Despite the differences, the step responses of both the vehicle (Figure 5) and the simulation follow the same behaviour. Figures 7 and 8 show the simulation velocity responses to brake step and throttle power-off inputs. The simulation responses match the vehicle responses (Figure 5) in terms of behaviour.

3. Controller Design

The outputs of the longitudinal control problem are i) the throttle angle, which controls the fuel/air mixture for the combustion process within the engine and ii) the brake pedal position, which applies a braking torque to each wheel. In Figure 9, the vehicle model's velocity responses to 100% throttle step input and 50% throttle step input are charted. The responses can be characterized as a second order over-damped with a slight delay. By comparing the 50% response multiplied by a factor of two with the 100% response we see that the vehicle model's response with respect to the throttle is clearly non-linear. In Figure 10 the vehicle model's velocity response to 100% brake step input and 50% brake step input are shown. The responses show that during braking, Coulomb friction dominates the system. It is clear that the vehicle response to a 50% brake step input is not half of the 100% signal indicating that the modeled braking system is non-linear. Figure 8 shows the vehicle model's velocity response when the throttle is disengaged, and can be considered a step input from 1 to 0. The throttle power off resembles the brake system's response although more gradual. It demonstrates Coulomb friction as well and can be considered a nonlinear response.

To address each of these nonlinear responses, different control systems are required depending on the operating conditions. Our approach is to divide the control space into regions within which the behaviour of the plant approximates linearity. A patch-work of linear controllers would then be able to address the entire operating envelope. These

linear controllers would all have the same form, but their gains would differ depending on the operating conditions. This common linear controller along with its collection of gains is considered a form of *adaptive control* referred to as *gain scheduling* (Astrom and Wittenmark 1994). The difference in our implementation of gain scheduling, is that the tedious task of determining each gain is accomplished using a machine learning algorithm called *Monte Carlo ES* reinforcement learning.

3.1 Reinforcement Learning

Reinforcement learning (RL) is a machine learning approach where a software agent senses the *environment* through its *states s* and responds to it through its *actions a* under the control of a *policy*, $a = \pi(s)$. This policy is improved iteratively through its experiences with the environment through a *reinforcement learning algorithm* which in this paper is called *Monte Carlo ES* (Figure 11). The environment provides the agent with numerical feedback called a *reward* for the current state, $r = R(s)$. The environment also supplies the next state based on the current state and the actions taken using the *transition function* σ, $s' = \sigma(s, a)$. In this study, the transition function is provided by the vehicle model. The control problem is formulated into mathematical framework known as a finite Markov Decision Process (MDP) (Bellman 1957) by defining $\{s, a, \pi, R(s), \sigma(s,a)\}$. The key feature of an MDP is that to be considered Markov, its current state must be independent of previous states. This is so that for each visit to a state, the software agent is given a path independent reward. Subsequent actions will result in new states giving rise to different rewards.

The challenge of reinforcement learning is to determine the actions which result in the maximum reward for every possible state, this state to action mapping is called the optimal policy π^* or the controller. For the current state, actions that result in more favorable future states lead to higher rewards. The favorability of a certain action given the current state is known as the *Q*-Value. As an agent experiences its environment, it updates the *Q*-Value for each state-action (s, a) pair it visits according to its reinforcement learning algorithm. As it repeatedly visits every (s, a), it updates the policy so that the highest valued (s, a) will dominate. The optimal policy is reached when every state-action pair results in the highest reward possible; that is when the *Q*-Value function has been maximized. The convergence of this maximization process requires that all states and actions be visited infinitely in order for estimates the *Q*-value to reach their actual values. To ensure this convergence criterion, policies leading to π^* are ε-soft, meaning that there is a ε probability that a random action or exploration is selected. Therefore, all actions and states will be reached as $t \to \infty$. This process of policy improvement is referred to as a reinforcement learning algorithm. Specifically, Monte Carlo reinforcement learning algorithms improve the policy using the averaged sample returns experienced by the agent at the end of each episode (Sutton and Barto 1998).

The key to the process of improvement is the reward function which expresses the desirability of being in a current state. It is the method of communicating to the agent the task to be performed. The challenge of the designer is to be able to come up with a reward function that captures the essence of the task so that learning can be achieved.

3.2 Longitudinal Control

Simply stated, longitudinal control of a vehicle is to be able to follow another vehicle in traffic without colliding into it. That is, the controller must maintain a relative speed of zero with the vehicle ahead while maintaining a fixed distance behind the forward vehicle; this fixed distance will be referred to as Δx_i. During the process of control, the vehicle's relative speed, $V_{rel} = \dot{x}_{i-1} - \dot{x}_i$ and range, $X_{rel} = x_{i-1} - x_i$, to the vehicle ahead will provide feedback to the control system. Figure 12 shows how multiple vehicle's are linked to provide longitudinal control for multiple vehicles.

Figure 13 shows the design of the control system, two parallel control systems are used, one for throttle control, and one for brake control. These two throttle and brake controllers are a combination of a digital Proportional-Derivative (PD) controller for V_{rel}, and a digital Proportional-Integral (PI) controller for X_{rel}. The difference equation which provides the throttle/brake command m_n is shown below

$$m_n = m_{n-1} + k_{pv}(v_n - v_{n-1}) + \frac{k_{dv}}{\Delta}(v_n - 2v_{n-1} + v_{n-2}) + k_{px}(x_n - x_{n-1}) + k_{ix}\Delta T x_n \quad (22)$$

where n is the current iteration of the control cycle, v is V_{rel}, x is X_{rel},, and ΔT is the period of the control cycle. Moreover, k_{pv}, kd_v, k_{px}, and k_{ix} are gains that are functions of MDP state variables s_1, s_2, and s_3 as described in Table 1. This allows simultaneous regulation of the relative speed as well as the range while reducing the steady state range error through the integral control of the range. The results of both the throttle and brake controllers are fed into a logic element controlled by the gain K_{coast} which decides whether throttle control or brake control is to be used. In this paper, K_{coast} is set to 0.25; that is throttle values less than 0.25 utilize the braking system rather than coasting. The logic for this element is shown below

$$\begin{aligned} &\text{if } (throttle > 0) \\ &\quad cmd_{throttle} = throttle, cmd_{brake} = 0 \\ &\text{else if } (throttle < -K_{coast}) \\ &\quad cmd_{throttle} = 0, cmd_{brake} = 0 \\ &\text{else} \\ &\quad cmd_{throttle} = 0, cmd_{brake} = 0 \end{aligned} \quad (23)$$

For a given operating point, there are eight parameters or gains which must be provided in a lookup table or schedule. By formulating the control problem into a MDP, the gain schedule can be learned using reinforcement learning. The episode is defined as starting at the onset of a change in Vx_{i-1} and ending when $Vx_i = Vx_{i-1}$ or when Vx_{i-1} has been changed. This follows the logic that when a new velocity is required, a set of gains should be selected from the gain schedule and applied for the duration of that command. The goodness of a set of gains can therefore only be assessed once the command is complete, thus the MDP is episodic in nature and the Monte Carlo ES reinforcement learning algorithm described in Figure 11 is used to learn the gain schedule.

The choice in the selection of states lies in the nonlinear nature of the throttle plant. At different initial speeds the throttle responds differently. Therefore, the controller gains will differ from a given initial speed to a final speed. In addition, the distance required to achieve this acceleration/deceleration which is reflected in the change in vehicle spacing is also an independent variable for the gain schedule. These three parameters are used as states (Table 1). The actions are the eight values which represent the gains used in the digital control system (Table 2).

Table 1 States of the longitudinal MDP

State	*Description*	*Digitization Sets*
S_1	Vx_0: initial vehicle speed	{ 5, 10, 15, 20, 25, 30, 35, 40} m/s
S_2	Vx_{i-1}: target vehicle speed	{ 5, 10, 15, 20, 25, 30, 35, 40} m/s
S_3	$\Delta x_f - \Delta x_f$: change in vehicle spacing	{-100, -90, -80, …, 80, 90, 100} m

Table 2 Actions of the longitudinal MDP

Action	*Description*	*Digitization Sets*
A_1	K_p: Throttle Proportional Gain (x)	{0.1, 0.2, 0.3, … 9.9} $n\_s = 100$
A_2	K_i: Throttle Integral Gain (x)	{0.01, 0.02, 0.03, … 0.99} $n\_s = 100$
A_3	K_d: Throttle Proportional Gain (V)	{0.1, 0.2, 0.3, … 9.9} $n\_s = 100$
A_4	K_{i2}: Throttle Derivative Gain (V)	{0.01, 0.02, 0.03, … 0.99} $n\_s = 100$
A_5	K_p: Brake Proportional Gain (x)	{0.1, 0.2, 0.3, … 9.9} $n\_s = 100$
A_6	K_i: Brake Integral Gain (x)	{0.01, 0.02, 0.03, … 0.99} $n\_s = 100$
A_7	K_d: Brake Proportional Gain (V)	{0.1, 0.2, 0.3, … 9.9} $n\_s = 100$
A_8	K_{i2}: Brake Derivative Gain (V)	{0.01, 0.02, 0.03, … 0.99} $n\_s = 100$

The reward function which reflects the specification of the control problem is a discrete function of the feedback variables, the current normalized relative speed and normalized relative velocity of the vehicle and is expressed below.

$$R_{Total} = \Re_V(V_{rel}) + \Re_X(X_{rel}) \tag{24}$$

$$R_X(X_{rel1}) = \begin{cases} 1 & if\ X_{rel} \le 0.1 \\ -1 & if\ X_{rel} < 0 \end{cases} \qquad R_V(V_{rel1}) = \quad if\ |V_{rel}| < 0.1$$

For a given episode, the solution which maximizes the reward, or minimizes the X_{rel} and V_{rel} without colliding with the vehicle ahead ($X_{rel} < 0$) will be favoured. These favoured solutions will be explored to determine the optimal solution.

4. Reinforcement Learning Experiments

The *RL* experiments obtain an optimal policy π^* for the longitudinal control of the vehicle. An experiment consists of 300 episodes where $\varepsilon = 0.25$ of the ε-soft greedy policy for a particular combination of the three states. For each episode, the agent must follow another vehicle placed ahead of it which is travelling at a constant speed. Once the leading vehicle has reached the end of the test track, the episode is complete. The distance of the test track is dependent on the speed of the lead vehicle using the following equation.

$$x_{\max} = 1 + 0.2 v_{lead}) \times 000 \qquad m \tag{25}$$

During each step of an episode, a reward is generated (24), this reward is accumulated during the course of an episode to measure the controller's tracking performance using a particular set of actions. Since it is possible to collide with the vehicle ahead during an episode, it would be beneficial if the reward were averaged to reflect how far the vehicle

made it during the course of the episode. Therefore, the average reward for the course of the entire episode is provided by the following equation.

$$R_{avg} = \frac{\sum_{i=0}^{\hat{}}}{x_{\max} - \cdot_{final}} \tag{26}$$

Figure 14 shows the average reward as the agent progresses through the learning cycle for a particular state combination. The learning performance is similar for all combinations. One can observe the steady increase in the average reward which eventually reaches a plateau.

The learned optimal policy is a collection of eight four-dimensional discrete hyperspaces, one for each gain of the longitudinal controller; that is four for the throttle controller and four for the brake controller.

$$\pi = \begin{cases} k_{pv}^{Throttle}(v_f, v_0, \Delta x_f - \omega x_0) & k_{pv}^{Brake}(v_f, v_0, \Delta x_f - \omega x_0) \\ k_{dv}(v_f, v_0, \Delta x_f - \omega x_0) & k_{dv}(v_f, v_0, \Delta x_f - \omega x_0) \\ k_{px}(v_f, v_0, \Delta x_f - \omega x_0) & k_{px}(v_f, v_0, \Delta x_f - \omega x_0) \\ k_{ix}(v_f, v_0, \Delta x_f - \omega x_0) & k_{ix}(v_f, v_0, \Delta x_f - \omega x_0) \end{cases} \tag{27}$$

5. Controller Performance Experiments

These experiments demonstrate the tracking performance of the optimal policy at various operating points. Three control situations are shown which form the basis of platoon maneuvers which allow vehicles to enter or exit formations. The first of these, shown in Figure 15 is speed control. The vehicle must reach a final speed of 20 *m*/*s* while maintaining a separation distance to the vehicle ahead of 20 *m*. At an initial speed of 10 *m*/*s*, the vehicle immediately decelerates to create room to accelerate to the higher speed. At an initial speed of 30 *m*/*s* the vehicle shows a negative range which means the vehicle cannot maintain its separation distance as it slows down.

The second control situation is shown in Figures 16 is referred to as negative range control. The vehicle must move from an inter-vehicle space of 15 *m* to 5 *m* while maintaining a speed of 10, 20, and 30 *m*/*s*. The closing of the gap is accomplished within 500 *m*, with minimal velocity fluctuation and no overshoot. Figure 17 shows the third control situation, positive range control. The vehicle must move from an inter-vehicle space of 5m to 15m while maintaining a speed of 10, 20, and 30 *m*/*s*. In opening the gap, the vehicle's velocity fluctuates during the manoeuvre with some overshoot in range. These experiments represent the basis for platoon maneuvers which allow vehicles to enter into the new open space or to close the formation when a vehicle has left.

6. Multi-Vehicle Performance Experiments

These experiments show the operation of the control system within a five car formation or platoon. Five control situations have been chosen to demonstrate the range tracking performance of the optimal policy for each of the four following vehicles.

Figure 18 shows the results of a five car formation moving at a constant speed of 20 *m*/*s*. In the first experiment, the inter-vehicle spacing is set to 5 *m* between each car. At time t = 0 *s*, *Car 2* is instructed to open the space in front to 15 *m*. The results show *Car 2* overshooting the 15 *m* to roughly 22 *m*, in 35 *s* the car has reached a steady-state

separation of 15 *m*, the following cars reach the steady-state by 70 *s*. In the second, the inter-vehicle spacing is set to 15 m between each car. At time t = 0 *s*, *Car 2* is instructed to close the space in front to 5 m. The results show *Car 2* reaching 5 *m* in 10 *s* without overshoot; the following cars reach 5 *m* in 50 *s*.

Figure 19 shows the results of a five car formation trying to maintain constant spacing while accelerating or decelerating from 20 *m*/*s*. In the first experiment, the inter-vehicle spacing is set to 20 *m* between each car. At time t = 0 *s*, *Car 1* is to accelerate to 30 *m*/*s*. The results show a close tracking of the velocity with the presence of oscillations. The second experiment shows the deceleration of the vehicle to 10 *m*/*s* with an inter-vehicle spacing of 15 *m*. The tracking of the velocity is excellent. Figure 20 shows an emergency stop situation with a 15 *m* inter-vehicle spacing. The tracking of the velocity and is excellent with a very steep deceleration. All vehicles stop without colliding into the vehicle ahead.

7. Conclusions

In this paper the nonlinear nature of the vehicle dynamics is shown. Due to the nonlinearities present in the engine model, the transmission model, and the tire model a complex nonlinear model results. From this, we conclude that linearization of the longitudinal model may not be suitable for the entire operating range of the vehicle. The linear controllers resulting from using a simplified linear model of the vehicle dynamics in the design process may only be adequate for a particular operating point.

The use of a more accurate nonlinear vehicle dynamics model in the design process should result in better nonlinear control systems for longitudinal control. In this paper, an adaptive control system using gain scheduling is introduced whereby the gains are learned using reinforcement learning. Even with a simple reward function, it is possible for Monte Carlo reinforcement learning to converge upon an optimal policy within 300 episodes for a particular operating regime; therefore, the MDP properly describes the task to be learned.

When the learned optimal policies are combined to provide an adaptive control surface or a gain schedule, nonlinear control is achieved throughout the operating range. The performance of the controller at specific operating points shows accurate tracking of both velocity and position in most cases. When the adaptive controller is deployed in a multi-vehicle convoy or platoon, the tracking performance is less smooth. As the second car attempts to track the leader, slight oscillations result. This oscillation is passed to the following cars, but as we move farther in the formation, the oscillations decrease, implying stability. The performance of the adaptive controller in a multi-vehicle convoy or platoon shows promise and forms the basis of higher level platoon maneuvers.

Acknowledgement

This research is funded by the Auto21 Network of Centres of Excellence, an automotive research and development program focusing on issues relating to the automobile in the 21st century. AUTO21 is a member of the Networks of Centres of Excellence of Canada program. Web site: http://www.auto21.ca

References

Varaiya, P. (1993) 'Smart cars on smart roads: problems of control'. *IEEE Transactions on Automatic Control*. Vol 32, March.

Shladover S. E., Desoer C. A., Hedrick J. K., Tomizuka M., Walrand J., Zhang W.B., McMahon D. H., Peng H., Sheikholeslam S., and McKeown N. (1991) 'Automatic vehicle control developments in the PATH program'. *IEEE Transactions in Vehicular Technology*. Vol. 40, no. 1, pp. 114-130.

Hedrick J.K., Tomizuka M., Varaiya P. (1994) 'Control issues in automated highway systems'. *IEEE Control Systems Magazine*, Volume: 14, Issue: 6, Dec, pp 21-32.

Ioannu P.A. and Chien C.C. (1993) 'Autonomous Intelligent Cruise Control'. *IEEE Transactions on Vehicular Technology*, Vol 42, No. 4, Nov, pp 657-672.

Raza H. and Ioannou P. (1997) 'Vehicle following control design for automated highway systems'. *Proceedings of 1997 IEEE 47th Vehicular Technology Conference*, Phoenix, AZ, USA, Vol 2, pp 904-908.

Maciuca D.B., Hedrick, J.K. (1995) 'Advanced Nonlinear Brake System Control for Vehicle Platooning'. *Proceedings of the third European Control Conference (ECC 1995)*, Rome, Italy.

Swaroop D., Hedrick J.K. (1994) 'Direct Adaptive Longitudinal Control for Vehicle Platoons'. *IEEE Conference on Decision and Control*, December.

Rajamani R., Tan H.S., Law B.K., Zhang W.B. (2000) 'Demonstration of integrated longitudinal and lateral control for the operation of automated vehicles in platoons'. *IEEE Transactions on Control Systems Technology*. Vol 8, Issue 4, July, pp 695-708.

Kato, S., Tsugawa S., Tokuda, K., Matsui T. Fujii, H. (2002) 'Cooperative Driving of Automated Vehicles with Inter-vehicle Communications'. *IEEE Transactions on Intelligent Transportation Systems*. Volume: 3, Issue: 3, pp 155- 161.

Zhang J. and Ioannou P.A. (2005) 'Adaptive Vehicle Following Control System with Variable Time Headways'. *Proceedings of 44th IEEE Conference on Decision and Control and 2005 European Control Conference. CDC-ECC '05*. pp 3880 – 3885.

Khatir, M. E., Davison, E. J. (2006) 'A Decentralized Lateral-Longitudinal Controller for a Platoon of Vehicles Operating on a Plane'. *Proceedings of 2006 American Control Conference*, June 14-16, Minneapolis, Minnesota, USA.

Thorpe C., Jochem T., Pomerleau, D. (1997) 'The 1997 automated highway free agent demonstration'. *Proceedings of IEEE Conference on Intelligent Transportation System, 1997*. ITSC 97. Boston, MA, USA, pp 495-501.

Tan H.S., Rajamani R., Zhang W.B. (1998) 'Demonstration of an automated highway platoon system'. *Proceedings of the American Control Conference 1998*. Philadelphia, PA, USA, vol.3, pp 1823-1827.

Tsugawa S., Kato S., Matsui T., Naganawa H., Fujii, H. (2000) 'An architecture for cooperative driving of automated vehicles'. *Proceedings of 2000 IEEE Intelligent Transportation Systems*. Dearborn, MI, USA, pp 422-427.

Michaud, F., Lepage, P., Frenette, P., Létourneau, D., Gaubert, N. (2006), 'Coordinated maneuvering of automated vehicles in platoon'. *IEEE Transactions on Intelligent Transportation Systems*, Special Issue on Cooperative Intelligent Vehicles, 7(4):437-447.

Haug, E. J. (1990), 'Feasibility Study and Conceptual Design of a National Advanced Driving Simulator', NHTSA Contract DTNH22-89-07352, Report No. DOT-HS-807-597, March.

Salaani, M. K., Grygier, P. A., Heydinger, G. J. (2001) 'Model Validation of the 1998 Chevrolet Malibu for the National Advanced Driving Simulator'. March 2001, SAE Paper 2001-01-0141.

McMahon, D. H; Hedrick, J K. (1989) 'Longitudinal model development for automated roadway vehicles', Institute of Transportation Studies California Partners for Advanced Transit and Highways (PATH), University of California, Berkeley, USA.

Peng H. and Tomizuka M. (1991) 'Optimal Preview Control For Vehicle Lateral Guidance'. In Research Reports: Paper UCB-ITS-PRR-91-16, Institute of Transportation Studies California Partners for Advanced Transit and Highways (PATH), University of California, Berkeley, USA.

Pham H., Hedrick K., Tomizuka M. (1994) 'Combined lateral and longitudinal control of vehicles for IVHS'. *Proceedings of the 1994 American Control Conference*, Vol.2, pp1205 – 1206.

Pham H., Tomizuka M., Hedrick K. (1997) 'Integrated Maneuvering Control for Automated Highway Systems Based on a Magnetic Reference Sensing System'. Research Reports: UCB-ITS-PRR-97-28, Institute of Transportation Studies California Partners for Advanced Transit and Highways (PATH), University of California, Berkeley, USA.

Peng H., Zhang W.B., Arai A., Lin Y., Hessburg T., Devlin P., Tomizuka M., Shladover S. (1992) "Experimental Automatic Lateral Control System for an Automobile", In Research Reports: UCB-ITS-PRR-92-11, Institute of Transportation Studies California Partners for Advanced Transit and Highways (PATH), University of California, Berkeley, USA.

Astrom K. J., Wittenmark B. (1994), *Adaptive Control*, Addison-Wesley.

Bellman R. E. (1957) 'A Markov decision process'. *Journal of Mathematical Mech*., Vol 6 pp 679-684.

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

Figures:

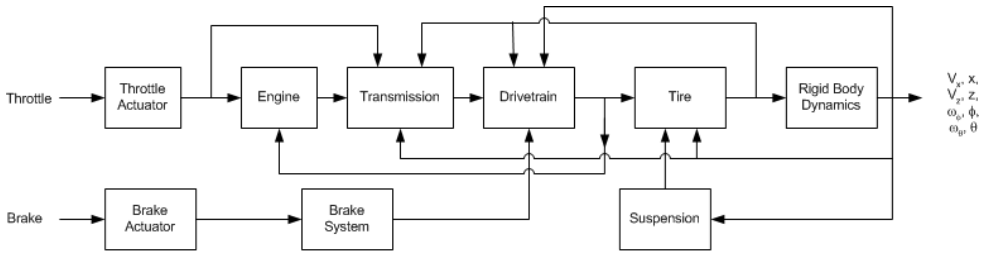

Figure 1 Schematic of the vehicle model

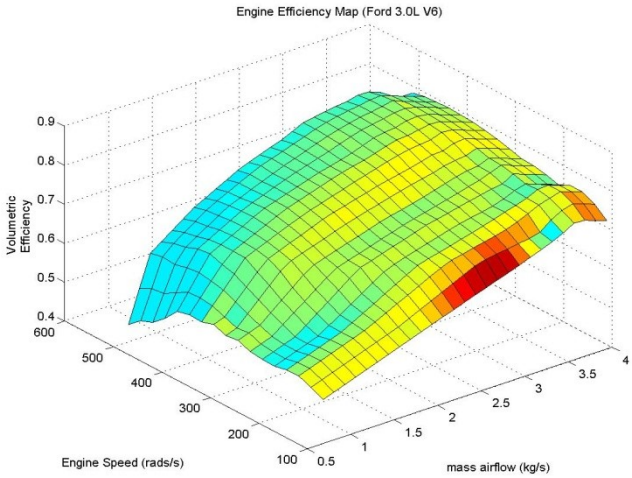

Figure 2 Engine volumetric efficiency surface

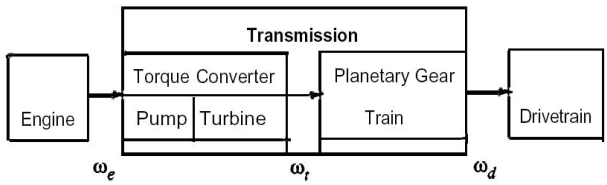

Figure 3 Schematic of transmission system (McMahon and Hedrick 1989)

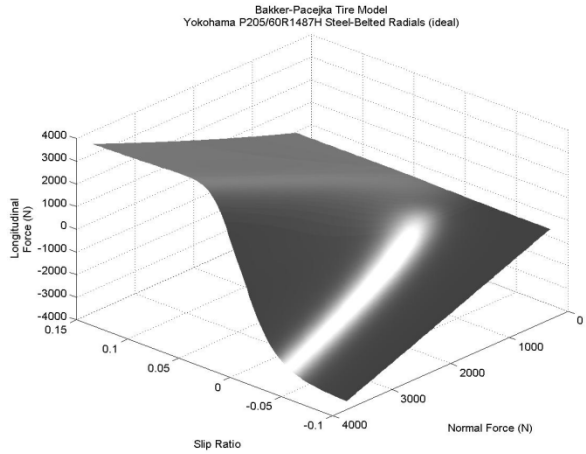

Figure 4 Longitudinal force-slip for *Yokohama P205/60R1487H* (ideal $\mu = 1.0$).

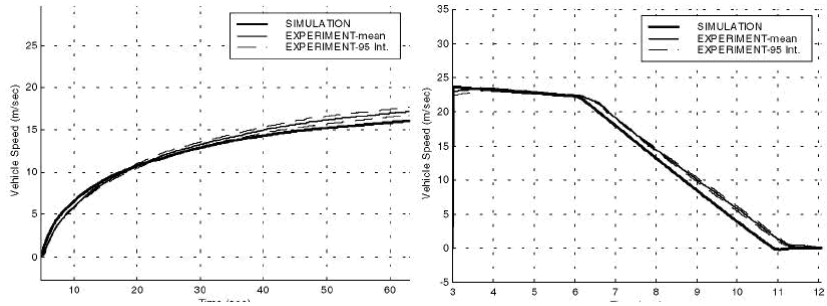

Figure 5 1997 Jeep Cherokee throttle step and brake step response (Salaani and Heydinger 2000)

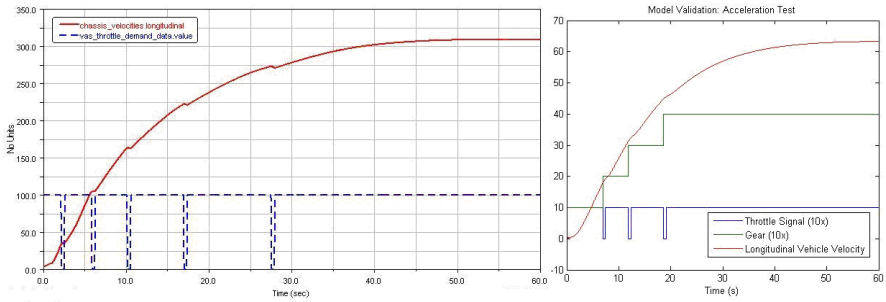

Figure 6 Adams Car and *Simulation* throttle step response

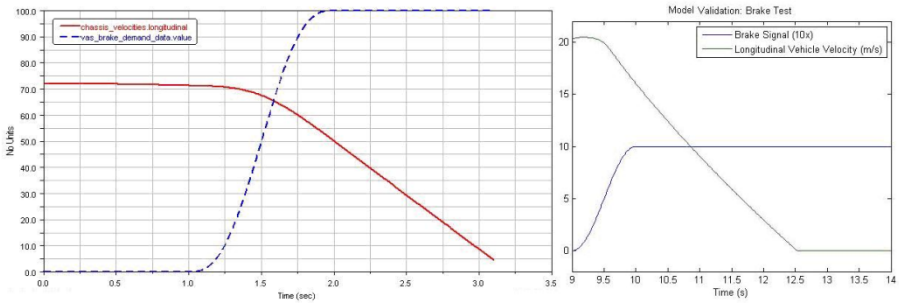

Figure 7 Adams Car and *Simulation* brake step response

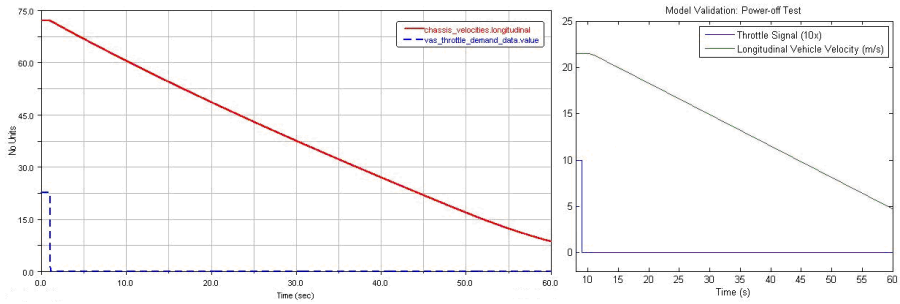

Figure 8 Adams Car and *Simulation* power-off response

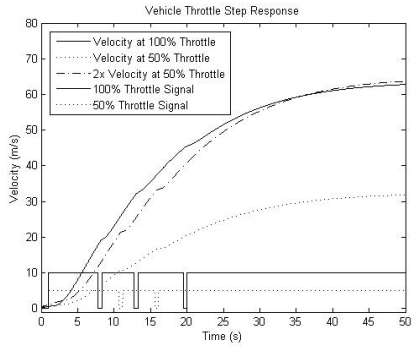

Figure 9 Vehicle model velocity responses to throttle step inputs

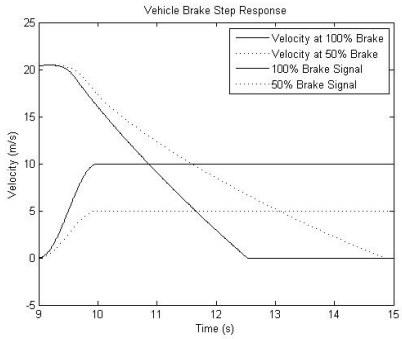

Figure 10 Vehicle model velocity responses to brake step inputs

Initialize, for all $s \in S$, $a \in A(s)$:
  $Q(s, a) \leftarrow$ arbitrary
  $\pi(s) \leftarrow$ arbitrary
  $Returns(s, a) \leftarrow$ empty list

Repeat forever:
  (a) Generate an episode using exploring starts π
  (b) For each pair (s,a) appearing in the episode
    $R \leftarrow$ return following the first occurrence of (s,a)
    Append R to $Returns(s,a)$
    $Q(s,a) \leftarrow$ average($Returns(s,a)$)
  (c) For each s in the episode:
    $\pi(s) \leftarrow \text{argmax}_a\, Q(s,a)$

Figure 11 Monte Carlo ES-algorithm

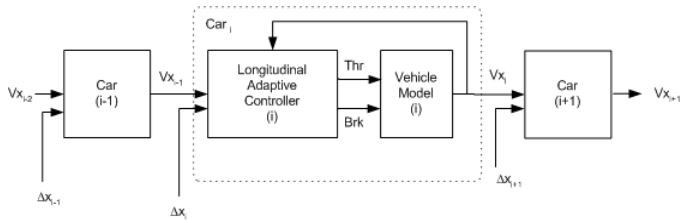

Figure 12 Overview of longitudinal control system

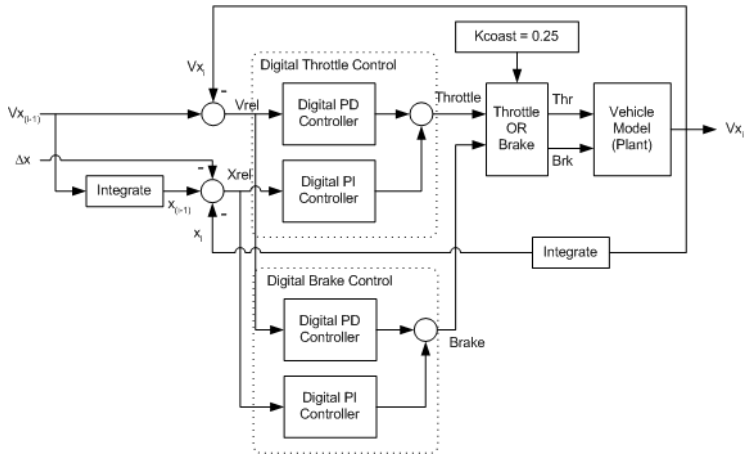

Figure 13 Block diagram of longitudinal controller

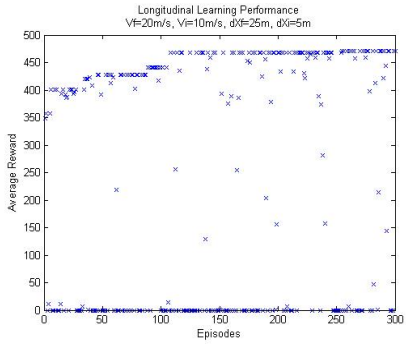

Figure 14 Performance of *Reinforcement Learning* experiments

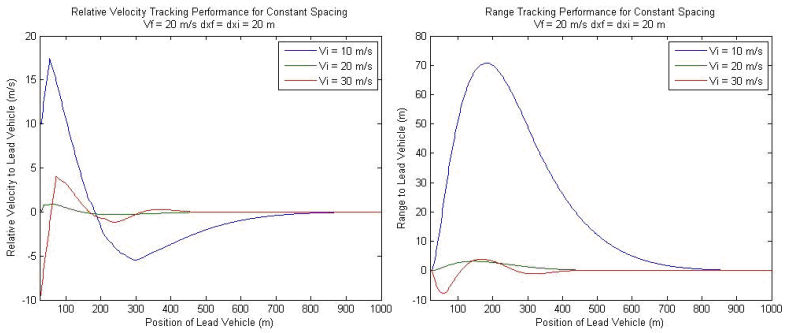

Figure 15 Speed control experiment

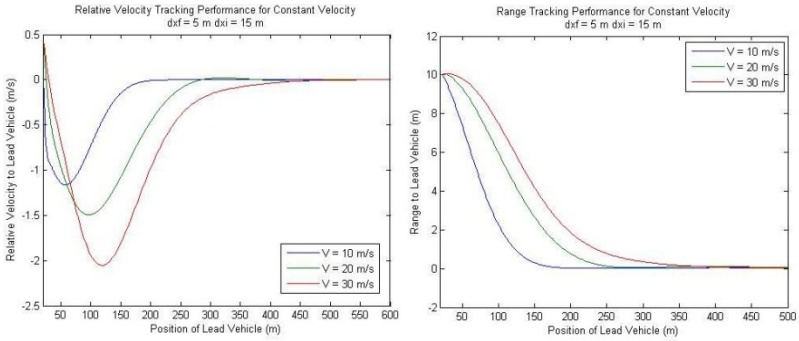

Figure 16 Negative range control experiment

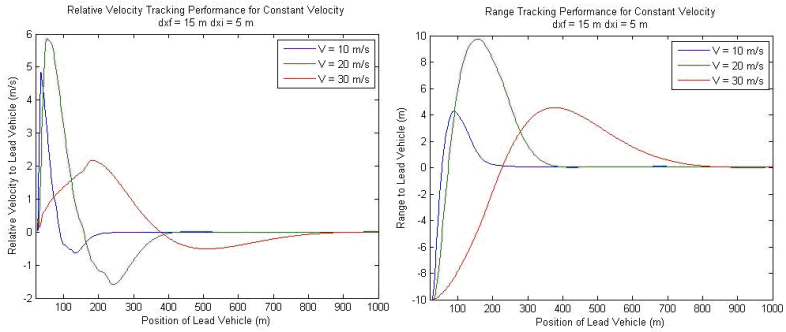

Figure 17 Positive range control experiment

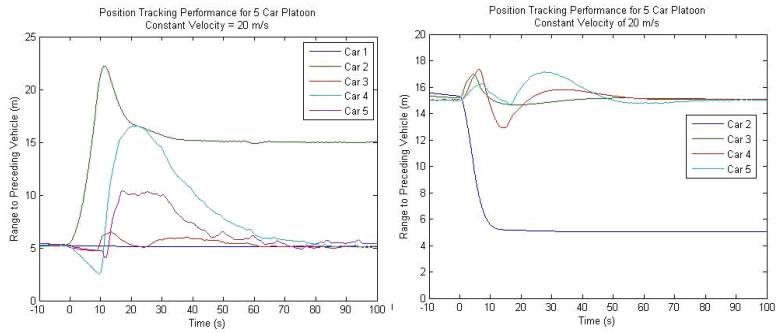

Figure 18 Multi-vehicle range control experiment: Open and Close

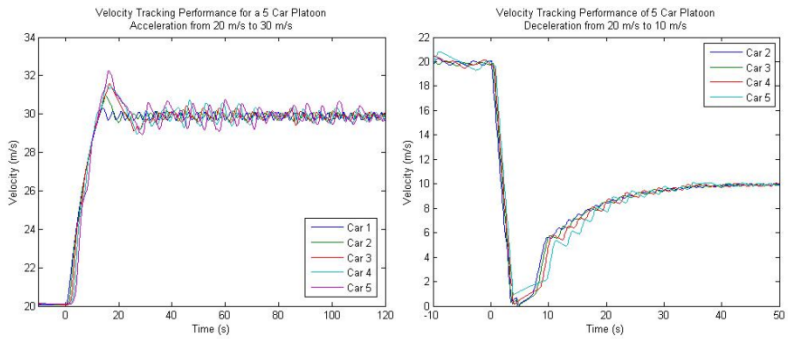

Figure 19 Multi-vehicle velocity control experiment: Acceleration and Deceleration

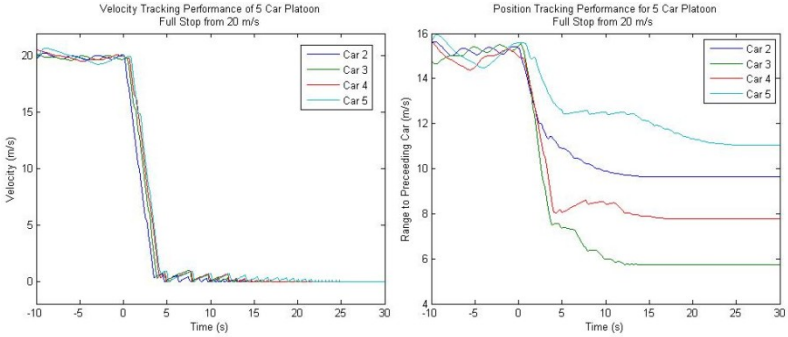

Figure 20 Multi-vehicle velocity control experiment (Emergency Stop)