Reinforcement Learning for Determining Temperature/Strain Behavior of Shape Memory Alloys

Kenton Kirkpatrick* and John Valasek†
Texas A&M University, College Station, Texas 77843-3141

The ability to actively control the shape of aerospace structures has led to the implementation of Shape Memory Alloy actuators. These actuators can be used for morphing or shape control by modulating their temperature, which is effectively done by applying a voltage difference across their length. Characterization of this temperature-strain relationship is currently done using constitutive models, which is time and labor intensive. Shape Memory Alloys also contain both major and minor hysteresis loops. Understanding the hysteresis is crucial for practical applications, and characterization of the minor hysteresis loops, which map the behavior of a wire that is not fully actuated, is not possible using the constitutive method. Numerical simulation using Reinforcement Learning has been used for determining the temperature–strain relationship and characterizing the major and minor hysteresis loops, and determining a control policy relating applied temperature changes to desired strain. This paper extends and improves upon the numerical simulation results, using an experimental hardware apparatus and improved Reinforcement Learning algorithms. Results presented in the paper verify the numerical simulation results for determining the temperature-strain major hysteresis loop behavior, and also determine the relationships of the minor hysteresis loops.

I. Introduction

Advancement of aerospace structures has led to an era where researchers now look to nature for ideas that will increase performance in aerospace vehicles. The main focus of the Texas Institute for Intelligent Bio-Nano Materials and Structures for Aerospace Vehicles is to revolutionize aircraft and space systems by advancing the research and development of biological and nano-technology.1 Birds have the natural ability to move their wings to adjust to different configurations of optimal performance. The ability for an aircraft to change its shape during flight in order to optimize its performance under different flight conditions and maneuvers would be revolutionary to the aerospace industry. To achieve the ability to morph an aircraft, exploration in the materials field has led to the idea of using Shape Memory Alloys (SMA) as actuators to drive the shape change of a wing. The most commonly used SMAs are composed of either nickel and titanium, or the combination of nickel, titanium, and copper. The benefits of using each of these alloys have been explored in this research.

*Graduate Research Assistant, Vehicle Systems & Control Laboratory, Aerospace Engineering Department. Student Member AIAA. kentonkirk@gmail.com
†Associate Professor and Director, Vehicle Systems & Control Laboratory, Aerospace Engineering Department. Associate Fellow AIAA. valasek@tamu.edu, Website: http://jungfrau.tamu.edu/valasek

American Institute of Aeronautics and Astronautics
SMAs have a unique ability known as the Shape Memory Effect (SME). This material can be put under a stress that leads to a plastic deformation and then fully recover to its original shape after heating it to a high temperature. This makes SMAs ideal for structures that undergo large amounts of stress, such as aircraft. SMAs begin in a crystalline structure of martensite and undergo a phase change to austenite as the alloy is heated. This phase transformation realigns the molecules so that the alloy returns to its original shape. This original shape is retained when the SMA is cooled back to a martensitic state, recovering the SMA from the strain that it had endured.

When a SMA wire has a phase transformation, it changes its length. The phase transformation from martensite to austenite causes a decrease in length while the reverse process extends it back to its original length. Control of this transformation is needed in order for morphing actuation to be possible. The SMA wire exhibits a hysteresis behavior in its relationship between temperature and strain due to non-uniformity in the phase transformations. This occurs because the phase transformation from martensite to austenite begins and ends at different temperatures than the reverse process. Figure 1 demonstrates this behavior, where $M_s$ is martensitic start, $M_f$ is martensitic finish, $A_s$ is austenitic start, and $A_f$ is austenitic finish.

\[
\rho c V \frac{dT}{dt} = \frac{V^2(t)}{R} - h A \left( T(t) - T_c \right)
\]

In Equation (1), $\rho$ is the wire density, $c$ is the specific heat of the wire, $V$ is the wire volume, $T$ is the wire temperature, $V$ is the voltage difference in the wire, $R$ is the wire electrical resistance, $h$ is the convective heat transfer coefficient, $A$ is the wire surface area, and $T_c$ is the ambient temperature of the coolant surrounding the wire.

This hysteresis behavior is most often characterized through the use of constitutive models that are based on material parameters or by models resulting from system identification. This is a time and labor intensive process that requires external supervision and does not actively discover the hysteresis in real-time. Other methods that characterize this behavior are phenomenological models, micromechanical models, and empirical models based on system identification. These models are quite accurate, but some only work for particular types of SMAs and most require complex computations. Many of them are also unable to be used in dynamic loading conditions.
making them unusable in the case of morphing. A major drawback to using any of these methods is that the minor hysteresis loops that correspond to a SMA that is not fully actuated are unattainable and must be determined through mathematical models. A simulated model of the major and minor hysteresis loops for an SMA wire is shown in Figure 2.

![Fig. 2 Results of SMA Hysteresis Simulation](image)

In order to map this SMA hysteresis autonomously and in real-time, this research uses the approach of an artificial intelligence program known as Adaptive-Reinforcement Learning Control (A-RLC). This is a form of machine learning that utilizes the interaction with multiple situations many times in order to discover the optimal path that must be taken to reach the pre-determined goal. By mapping the behavior in real-time, both major and minor hysteresis loops can be experimentally determined.

A-RLC is ideal for the morphing technology because it allows for a machine to learn the optimal control in real-time with no external supervision.\textsuperscript{11} By creating and updating a control policy based on states, actions, and goals, A-RLC can begin by exploration and move to exploitation once it finds the optimal actions to take for each individual state. A-RLC uses a process of rewards and consequences that allow the program to remember which actions are good at reaching the goal and which are poor. Reinforcement Learning (RL) is the driving force of SMA hysteresis learning that will eventually lead to SMA shape control.\textsuperscript{11,12,13,14,15} The SMA phase transformation is not a thermodynamically reversible process. This leaves uncertainty in the model due to the highly non-linear behavior of the SMA. This model is unknown and needs to be determined by the RL in conjunction with the experimental setup. The control policy is also unknown initially so RL is exploited because it does not require a pre-defined control policy.

This paper follows the development of an RL algorithm that can actively learn the hysteresis behavior in a SMA wire. RL is used to determine the major and minor hysteresis behavior in an SMA wire, and the algorithm is validated using an experimental hardware apparatus for the training, testing, and experimentation of specimen SMA wires. Details of the hardware/software interface for real-time experimentation are provided, and results are verified by comparison to constitutive and mathematical models.

### II. Reinforcement Learning

Reinforcement Learning is a process of learning through interaction in which a program uses previous knowledge of the results of its actions in each situation to make an informed decision when it later returns to the same situation. RL uses a control policy that is a function of the states and actions. This control policy is essentially a large matrix
that is composed of every possible state for the rows, and every possible action for the columns. In this research, a third dimension is included in the control policy that is composed of every possible goal state.

One of the most commonly implemented forms of RL is Temporal Difference. Temporal Difference methods have the advantage of being able to learn at every time step without requiring the input of an environmental model. This research utilizes a method of Temporal Difference known as Sarsa. Sarsa is an on-policy form of Temporal Difference, meaning that at every time interval the control policy is evaluated and improved. Sarsa updates the control policy by using the current state, current action, future reward, future state, and future action to dictate the transition from one state/action pair to the next. The action value function used to update this control policy is:

$$Q_k(s, a) = Q_k(s, a) + \alpha [r_k + \gamma Q_{k+1}(s', a') - Q_k(s, a)]$$  \hspace{1cm} (2)

where $s$ is the current state, $a$ is the current action, $Q$ is the control policy, and the $k$ term signifies the current policy. The $\alpha$ term is a parameter that is used to keep the RL from repeating itself within each episode. The term $s'$ refers to the future state, $a'$ is the future action, $k+1$ corresponds to the future policy, and $\gamma$ represents a constant that is used to optimize the rate of convergence by weighting the future policy. The reward given for each state/action pair is defined by $r$. The reward that is given for each situation is a user-defined parameter. For this research, when a goal state is achieved, a reward of 1 is given, while a reward of 0 is given for any other state within range. If the boundaries of the problem are exceeded, a reward of -1 is given to discourage following that path again. In this research, the control policy was modified to a three-dimensional matrix that includes the goal as the third dimension. With $g$ representing the goal state, the action value function now becomes:

$$Q_k(s, a, g) = Q_k(s, a, g) + \alpha [r_k + \gamma Q_{k+1}(s', a', g) - Q_k(s, a, g)]$$  \hspace{1cm} (3)

This action value function creates the policy that can be used to learn the parameters of the system being explored through RL. The Sarsa method uses a simple algorithm to update the policy using the action value function provided in Equation (3). This algorithm involves simply choosing the action from the policy, $Q$, based on what the expected reward is and updating the policy using Equation (3) based on what reward is received.

The $\varepsilon$-Greedy method of choosing an action is used in this research, which means that for some percentage of the time that an action is chosen, the RL will choose to randomly explore rather than choose the action that the control policy declares is the best. This is because the RL might not have already explored every possible option, and a better path may exist than the one that is presently thought to yield the greatest reward. A fully greedy method chooses only the optimal path without ever choosing to explore new paths, which corresponds to an $\varepsilon$-Greedy method where $\varepsilon = 0$. A random number generator is used to determine whether exploration or exploitation will be used at each action choice, and the value of $\varepsilon$ determines the probability of exploration occurring. For this research, a progressively changing $\varepsilon$-Greedy method was implemented by altering the exploration constant, $\varepsilon$, depending upon the current episode.

In this research, the states are defined by the current strain and temperature, while the actions are defined by the desired temperature. The desired temperature is immediately converted to voltage that is applied to the SMA wire. The goal that the system is attempting to reach is the desired strain of the SMA wire. The purpose of the RL agent is
to converge on the optimal temperature needed to produce the desired strain based on the current strain in the wire and the temperature at which the wire currently exist.

The RL method described up to this point will solve a problem where discrete values of strain states and temperature states are needed, but the SMA wire used in this experimentation is a physical specimen that has a continuous state-space. Its action-value function $Q(s,a)$ and action-preference function $p_w(s)$ are both defined on continuous domains. One approach to solving this type of Reinforcement Learning problem with a continuous state-space is function approximation methods. For this research, a 1-Nearest Neighbor approach was implemented. By using this formation of k-Nearest Neighbor, if the current state falls between two of the discrete strain/temperature values in the state space, the values for the closest state are used.

Once the RL learns the optimal temperature required to achieve each goal strain from each initial strain, it can then be used to map the hysteresis behavior of the SMA wire in real-time. By allowing RL to run through each of its learned situations and recording the strain and temperature data at each time interval, the characterized hysteresis loop can be easily plotted to graphically show that it has learned the SMA phase transformation strain/temperature behavior.

III. Experimental Apparatus

For the SMA wire to be tested, a physical experimental setup was first constructed. The SMA exists in an apparatus that is constructed of Plexiglas and aluminum supports. The apparatus is sealed so that no fluid can leak out as the experimentation is proceeding. The wire is attached to the walls by Kevlar chords and is set in series with a spring with constant $k = 4.34$N/mm. It is tightened so that the wire experiences a stress of approximately 100MPa in its initial martensitic state at zero strain.

A thermocouple is connected to the wire, which measures the temperature of the wire and sends small voltages to the Data Acquisition (DAQ) board. A Linear Voltage Differential Transducer (LVDT) is supported above the fluid by an aluminum beam, and the probe end is connected to the Kevlar chords for position measurement without receiving current from the wire. The LVDT sends a voltage to the DAQ board which changes depending on the position of the probe. A variable voltage supply is used to provide a voltage difference across the wire for heating it and is connected to the SMA wire via alligator clips. The voltage supply receives its commands from the DAQ board with an input/output voltage ratio of 3.6 and outputs voltages in the range of 0.00V-2.50V. Figure 3 shows the complete experimental apparatus.

![Fig. 3 Experimental Apparatus](image-url)
The apparatus contains a pool of antifreeze which completely submerges the SMA wire and the alligator clips to allow sufficient cooling of the wire for prevention of overheating and to decrease the time required for the reverse phase transformation from austenite to martensite. The antifreeze is drawn out of the apparatus by a pump that sends it into a pool for temperature regulation. The pool contains both heating and cooling coils that allow it to keep the antifreeze at a specified ambient temperature. In this research, the ambient temperature is kept at 21°C. The cooled antifreeze is then drawn back out of the temperature regulation pool by another pump and is sent into the apparatus to continue fluid circulation and keep the coolant at a constant room temperature.

The setup of this hardware led to many technical issues, some of which revealed key conclusions about the ability to characterize an SMA wire using this real-time method. The coolant originally used for decreasing the time required to achieve martensite was water. Water was assumed to be a good fluid to use as it was readily available and has low electrical conductivity. Temperature regulation for water is also very easy, making it an obvious choice for the coolant. However, this research has revealed that water was not an ideal coolant for this particular case. Water transfers heat too easily, leading to poor temperature measurements by the thermocouple. This occurs due to the fact that the thermocouple experiences large temperature differences between the water touching the wire and the water at ambient temperature. In addition, water cannot exceed 100°C while in its liquid state so temperature measurements at high temperatures become highly inaccurate and useless for application in this experiment. The water also causes some current loss due to impurities in the water so high voltages (10-12V) are required to achieve high temperatures. The characterization of the major hysteresis loop using direct user input for a water-filled apparatus is shown in Figure 4.

By using ethylene glycol (antifreeze) as a coolant instead of water, these problems can be overcome. Antifreeze does not transfer heat as easily as water so the ambient temperature in the apparatus does not affect the antifreeze that touches the SMA wire as quickly. This allows for much smoother temperature measurements throughout the phase transformations, although it does cause a slower phase transformation back to martensite. Antifreeze also has the ability to greatly exceed the previous limit of 100°C without boiling, thereby eliminating the turbulence effects caused by water at high temperatures and allowing for better measurements. Antifreeze is a very good insulator, and by using antifreeze, full actuation can occur with 2.5V instead of the 12V required in water. The characterization of the major hysteresis behavior using direct user input in an antifreeze-filled apparatus is shown in Figure 5.
In Figure 5, the experimental results are compared to the mathematical model that was used in the simulation portion of SMA characterization. This model is based on a hyperbolic tangent curve that is represented by Equations (4) and (5):

\[
M_i = H / 2 \tanh \left( \left( T - ct_i \right) a \right) + s \left( T - (ct_i + ct_s)/2 \right) + H / 2 + cs
\]

\[
M_r = H / 2 \tanh \left( \left( T - ct_r \right) a \right) + s \left( T - (ct_r + ct_s)/2 \right) + H / 2 + cs
\]

In these equations, \( H \), \( c_t \), \( a \), \( s \), \( c_t \), and \( cs \) are constants that determine the shape of the hyperbolic tangent model. \( M_i \) and \( M_r \) are the strain values that correspond to the temperature input into the equations. The constants were selected by creating a curve that best fit a known hysteresis behavior for a SMA wire.

**IV. Results**

This research initially used a CuNiTi wire for testing, which has the favorable property of taking much more stress to fail than a NiTi SMA wire allows. However, this research has uncovered issues with using this type of SMA because of poor hysteresis characterization over a period of a few episodes. Figure 7 shows a plot of the hysteresis behavior of a CuNiTi wire as obtained over a course of 3 episodes.

As can be seen in Figure 7, the hysteresis behavior does not appear nearly as clearly as it does with the NiTi wire shown in Figure 5. Due to this fact, the CuNiTi wire was replaced by a NiTi wire for the remainder of the tests. By using the NiTi wire, the lower tensile strength became a problem. The spring that was used to keep the wire under constant stress was replaced with a dead weight providing a tensile stress of 105 MPa. The dead weight is a superior method of providing stress in this case because it provides a constant stress that does not increase with SMA strain.
Due to the use of dead weight, the NiTi wire does not break as easily as before, allowing data to be recorded using the same sample for a larger period of time.

This experiment has been tested over many episodes at several different goal states corresponding to individual strain states, where an episode is defined as the achievement of a goal. With the current configuration, 3.3% strain is the maximum strain possible that corresponds to complete actuation. In order to demonstrate the convergence of the RL program, a goal state of 2.7% was investigated in detail. This goal was chosen because it requires nearly complete actuation of the SMA wire, but does not reach a fully actuated state. This forces the RL program to find the correct temperature state exactly. When the maximum goal state of 3.3% is chosen the state is achieved more easily since any temperature exceeding the austenite finish temperature will yield a fully actuated strain state. This makes observing an intermediate strain state much more useful.

Figure 8 shows the relationship between the episodes completed and the total Reinforcement Learning actions attempted in order to reach a goal of 2.7% strain. Every episode presented in this data begins at a fully un-actuated strain of 0%. As this graph shows, the RL algorithm takes fewer actions to achieve the desired goal state as it experiences more episodes. This proves that the RL becomes more successful in completing its objective of finding the optimal temperature required to achieve this goal state as it continues to learn.

![Fig. 8 Episodes vs. Actions](image)

Figure 8 reveals that the control policy begins learning enough about the system to obtain the desired strain with only a few actions by the time it has reached 20 to 25 episodes. However, it can also be seen that even after this point there are a few episodes that required a larger number of actions in order to find the goal. This happens because the RL algorithm being used incorporates the logic of the ε-Greedy method. Even after the algorithm begins converging on the optimal policy, exploration is still encouraged in order to allow the system to find a better path to goal state achievement.

Over the course of 37 episodes to a goal state of 2.7% strain and back to a goal state of 0% strain, the major hysteresis behavior becomes visible. Figure 9 shows that the major hysteresis behavior is experimentally attainable from Reinforcement Learning.
The progression of the control policy’s ability to obtain the hysteresis behavior is also of interest from this experiment. This information shows how well the experiment is able to utilize the learning capabilities of a RL algorithm. Figure 10 shows the paths that are taken to obtain the final goal state for three different episodes that are represented in the convergence behavior shown in Figure 8.

During episode 12, the experimental system required 147 actions in order to achieve the goal strain of 2.7%. As a result, the system wandered between many different temperatures before it was finally able to find the temperature that would yield the correct goal state. After running more similar episodes, the control policy learned how to achieve the goal state while taking fewer actions. By episode 23, only 4 actions were required to achieve the goal of 2.7% strain. Episode 30 demonstrates the control policy’s ability to find the correct goal state in only 1 action. Figure 10 shows the affects of the RL algorithm’s convergence upon an optimal control policy.

Reinforcement Learning’s ability to find a control policy that learns the minor hysteresis behavior of a Shape Memory Alloy is of special interest because minor hysteresis loops are difficult to obtain by other methods. By using RL to characterize the hysteresis behavior, the minor loops are obtained just as easily as the major loops. The minor hysteresis behavior can be extracted from individual episodes, as is demonstrated in Figure 11.
Figure 11 represents the extraction of the major hysteresis loop and 4 minor hysteresis loops from episode 12 of the 2.7% goal experimentation. Normally these minor loops must be obtained by using mathematical models based on the major hysteresis behavior, but this shows that the minor hysteresis loops can be experimentally obtained through the RL method. The orange loops show those which start and end at an either fully unactuated or fully actuated SMA. The red loop reveals that not only those are possible, but characterizations of loops that are completely within the interior are possible as well. The real-time data collection as the RL algorithm experimentally determines how to achieve each goal state allows both major and minor hysteresis loops to be mapped precisely.

V. Conclusions and Future Research

This research has made several conclusions about the characterization of SMA wires using Reinforcement Learning. It has been determined that water is a poor coolant for this approach, while antifreeze provides a remedy to the problems presented by water. The experimentation using both a spring and a dead weight as stress methods has revealed that a spring is a poor tool for use in RL control. The dead weight provides a system with fewer variables and allows experimentation with a NiTi SMA specimen. This experiment also concluded that NiTi wires are superior to CuNiTi wires due to the fact that the hysteresis behavior is less extreme and more difficult to model in a copper-based wire. It was also concluded that the Reinforcement Learning approach does indeed accomplish its goal of converging on the optimal temperature for achieving a particular goal state, which allows the program to learn the recorded temperature and strain data required to characterize the hysteresis behavior. This will allow the future goal of controlling the length of SMA wires to be accomplished.

The current progress of this work involves using a similar algorithm to learn the voltage/position behavior of a SMA wire. In order for this to be useful in the morphing of aerospace structures, the length of the wire needs to be controlled. Since the voltage needed to create a desired strain can not currently be intuitively drawn from the temperature/strain relationship, the next step is to use simulations to decipher how to use RL to control the length of the wire. This process is currently under way, and should be completed over the next few months. Once that is complete, this will be extended into an experimental setup that will use RL to actively control the length of a SMA wire in a hardware situation.

Acknowledgment

The material is based upon work supported by NASA under award no. NCC-1-02038. Support for this work has also been received from the National Science Foundation Graduate Research Fellowship Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Aeronautics and Space Administration or the National Science Foundation.

VI. References


