CHAPTER 2
LITERATURE REVIEW

2.1 - Introduction

A review of existing literature was performed to support the study undertaken in this thesis. A general survey was first performed to chronicle past research efforts in developing travel surveillance technologies used for ATIS purposes. Next, the potential advantages of AVI technology versus loop detection and GPS technologies are discussed. Lastly, the power of neural network technology was introduced to support the travel time forecasting studies in *Chapter 5*.

2.2 – Survey of Research in AVI and Other Traffic Surveillance Technologies

In the recent past, researchers have tested a wide array of technologies in an attempt to find improved methods of monitoring traffic conditions. Those techniques can be grouped into roadside techniques and vehicle techniques. Roadside techniques use detecting devices physically located along the study routes whereas vehicle techniques use detecting devices carried inside vehicles. AVI system comprises one of those advanced technologies currently be used. A brief survey of technologies explored during the past decade to provide an understanding of the level of research interest in traffic surveillance technologies.

Bohnke and Pfannerstill (1986) introduced a pattern recognition algorithm, which could utilize unique vehicle presence signatures generated by successive series of inductance loop detectors system. By identifying and reidentifying platoons of vehicles traveling across links bounded by loop detection equipment, vehicle travel times could be determined.

Ju and Maze (1989) performed simulations on incident detection strategies using the FREQ8PE simulation model. Their research evaluated a comparison of incident detection strategies using police patrol versus the use of motorist call boxes at 1-km spacing. The motorist call boxes formed the backbone of the modeled freeway surveillance and control system (FSCS). This FSCS yielded a benefit-to-cost ratio of 2.69 as it generated benefits from travel-time reduction and reduced fuel consumption. These benefits were brought about by reduced incident detection time afforded by the motorist call boxes.

In the development of video-based surveillance, Berka and Lall claim that loop detection reliability is low, and that maintenance and repair of such a pavement-based system creates safety risks for repair crews. Berka and Lall maintain that non-intrusive technologies such as video surveillance provides reduced traffic disruption during installation or repair. In addition, video surveillance is capable of detecting incidents on the sides of roadways, outside of the detection range of loop detectors.

Automatic vehicle identification (AVI) represents a major technological advance in the traffic surveillance technology (Bergan, Henion, et. al, 1987). It origins from the railway industry to monitor the movement of trains; to enable efficient scheduling; and, more importantly, to reduce potential conflicts or collisions. Prior to the installation of an AVI system in Houston, Texas, there already had several AVI system existed, which including Hong Kong Electronic Road-Pricing Project (1983); San Francisco International Airport toll revenue collection (1985); Singapore Road-Pricing study (1986); Heavy vehicle Electronic License Plate (HELP) program (1991) etc. All those project suggests that the implementation of accurate, dependable AVI system is currently possible and the use of AVI systems has the great potential to provide significant monetary saving.

In 1991, a cellular phone demonstration project was designed to monitor freeway traffic conditions in north Houston as a test of Houston AVI system. Researchers recruited 200 volunteers to participate in the program, which required them to call a traffic information office when they passed specific freeway locations during their morning and evening commutes. The lessons learned from the cell phone project aided in the development of the data analysis, processing and dissemination techniques used for the AVI system that was later constructed in Houston and San Antonio. In a similar scenario, prior to installing a large-scale AVI system in the Puget Sound area, a small-scale test of AVI was performed (Butterfield et. al, 1994). In this test, AVI was “piggy-backed” with existing loop detectors. Results yielded an AVI detection rate of about 80% for a fleet of tag-equipped buses.

In a 1996 report by Turner, a variety of techniques for travel time data collection were discussed, along with the advantages and disadvantages of each. These data collection techniques included electronic distance measuring instruments (DMI's), License plates matching, Cellular phone tracking, Automatic vehicle location (AVL), Automatic vehicle identification (AVI) and Video imaging. Turner specifically noted that travel time information was of particular importance for applications including congestion measurement and real-time travel information.

In this brief survey, more than ten distinct traffic surveillance technologies have been identified as the subject of research efforts since 1986. The amount of attention given to the research field of traffic surveillance clearly suggests that a surveillance system that can provide reliable and accurate travel time data would have great potential. The research community's interest in developing reliable and accurate surveillance systems is a primary motivation for the evaluation of San Antonio's AVI system.

2.3 - Potential Advantages of AVI over Inductance Loop Detection

The main advantage AVI offers over loop detection is its ability to provide *space mean speed* information, which is involved in the flow-density relationships of traffic studies. Loop detectors monitor traffic conditions at single-point locations where the detector is located. These loops are capable of generating *spot mean speed* data at various points along a traffic facility. The spot mean speeds must then be processed to estimate the speeds of vehicles between the detector location points. Given that loop detector spacing is often ½-mile or greater, there can be significant uncertainty in attempting to estimate the speed of vehicles between loops. Ford (1998) notes that inaccurate results can be generated by loops because they do not easily identify congestion that occurs between loop stations. He specifically reports that previous research has found loops to be inaccurate in both congested and high-speed conditions, with expected error measurements ranging from 5-10 mph. In addition, loop detectors are prone to failure.

AVI, meanwhile, can monitor the progress of vehicles across links of traffic, giving travel time information more accurate than that derived from loop detector measurements. Turner (1998) further attests to the advantages of AVI in acquiring travel time data. He reports that a study by NCHRP showed that travel time data generated from space mean speed measurements are rigorous enough for technical analyses while being simple enough to be understood by non-technical audiences.

In a comprehensive comparison of loop detection and AVI technologies for the collection of travel times, the assessments are given by author in Table 2.1.

Data Collection Technology	**Costs**			**Data Accuracy**	**Remark**
	Capital	**Installation**	**Data Collection**		
Loop Detector System	Low	Moderate	Low	Low	High failure rate and inaccurate estimations
AVI Systems	High	High	Low	High	Limited to fixed route and checkpoints and probe density

Table 2.1 – AVI System VS Loop Detector System

In spite of higher costs, Ford concludes that the reliability and accuracy of AVI equipment makes it a better option than loop detection for Advanced Traveler Information Systems (ATIS) and whole intelligent transportation system (ITS).

Parkany & Bernstein (1995) discuss the potential advantages that vehicle-to-roadside communications (VRC) such as AVI could have in incident detection. Compared to spot speed, occupancy and flow data provided by loop detectors, AVI could provide transportation officials with more useful traffic data. This data can include lane-specific and station-specific headways, the volume of tag-equipped vehicles on a section of a facility at any time, and the number of tagged vehicles that switch lanes between readers. Preliminary conclusions from Parkany and Bernstein’s research indicated that a VRC incident detection system using headway, lane-switch and lane-monitoring algorithms could perform better than the California algorithm typically used with loop detectors.

2.4 - Potential Advantages of AVI over GPS

In addition to its advantages over loop detector surveillance, AVI also has a distinct advantage over the use of GPS. Moore (1999) presents the case of downtown “canyons” created by skyscrapers. In such cases, GPS communications can be blocked by the presence of large buildings. In these downtown locations where improved efficiency is often needed, terrestrial-based AVI systems can perform very capably.

In recent years, researchers have also been looking to AVI to help improve incident detection on freeways and arterials. Historically, algorithms which analyze loop detector data have performed incident detection. Ivan and Chen (1997) compared several algorithms using both fixed and vehicle-based surveillance methods. Their results indicated that a combination of the two types of surveillance methods yielded the best incident detection results.

Petty et al. (1997) concluded that a probe-vehicle-based algorithm for incident detection is feasible, and it avoids certain infrastructure-related problems facing loop-based algorithms. Similarly, Marshall and Batz (1994) noted that AVI equipment used in the electronic toll and traffic management (ETTM) system constructed in the Greater New York/New Jersey Metropolitan Area offered more reliable potential for incident detection. This reliability stemmed from the individual vehicle travel times gathered by the system.

In a comprehensive comparison of GPS and AVI technologies for the collection of travel times, the assessments was given by author in Table 2.2.

Data Collection Technology	**Costs**			**Data Accuracy**	**Remark**
	Capital	**Installation**	**Data Collection**		
GPS Technology	Very High	Very high	Low	Very High	High failure rate in downtown area since “canyons”
AVI Systems	High	High	Low	High	Limited to fixed route and checkpoints and probe density

Table 2.2 – AVI System VS GPS Technology

In addition to AVI's reported ability to provide reliable travel time information and potential to improve incident detection methods, AVI offers more flexibility in its potential uses in transportation management programs than traditional loop detection systems. Inherent to the use of AVI technology is its ability to track individual vehicles, a capability that loop detection does not possess. Turner mentions that, in addition to its ability to provide real time travel information, AVI is even more valuable because of its use in electronic toll collection and fleet management applications. In addition, travel-time information is fast becoming an integral part of real-time travel information systems used in such applications as in-vehicle navigation.

Dorrance notes that the real-time data provided by AVI can be used to evaluate the effects of traffic management strategies as well as to help develop new management programs. In addition, such information can be used to market HOV lanes, given that comparisons of HOV-lane speeds vs. non-HOV-lane speeds can be provided to travelers. By relaying such up-to-the-second speed data to commuters, HOV lanes could become more attractive to the potential traveler if he receives a quantifiable report indicating that HOV traffic is moving faster than non-HOV traffic. Such information could lead to more drivers deciding to car pool, thus reducing the amount of traffic on the road. Incident assessment, emergency vehicle routing and traffic flow pattern monitoring are other potential advantages of AVI cited by Dorrance. As a final note, Levin and McCasland report that AVI programs hold unique potential for improving relationships between transportation management officials and the general public. The use of travel tags provides a physical connection between the two groups, which can help foster support for future traffic management projects.

Without question, AVI holds substantial promise in improving upon traffic surveillance capabilities currently offered by inductance loop detection technology as well as GPS technologies.

2.5 – The Power of Neural Networks

Neural networks, or simply neural nets, are computing systems, which can be trained to learn a complex relationship between two, or many variables or data sets. Basically, they are parallel computing systems composed of interconnecting simple processing nodes (Lau, 1992). Neural net techniques have been successfully applied in various fields such as function approximation, traffic flow prediction, waterway lock service times, marketing forecasting, incident detection and signal control. Some examples that directly apply to transportation are truck brake diagnosis systems, vehicle scheduling, and routing systems. In the present application, a fully connected multilayer feedforward neural network combined with a backpropagation algorithm will be used to forecast link travel times using San Antonio AVI tag data.

Neural networks utilize a matrix programming environment making most nets mathematically challenging. It should be understood that it is only the intent here to give the reader a brief synopsis of neural networks, and describe the basic type of neural network used for the research. For a more in-depth review of the intensive mathematical derivation and computation of the neural networks please refer to the listed references mentioned throughout this section.

2.5.1 - Elements of the Neural Network Paradigm (Hagan *et al.* 1996)

The neuron model and the architecture of a neural network describe how a network transforms its input into output. This transformation can be viewed as a computation. The model and the architecture each place limitations on what a particular neural net can compute.

Each neuron is represented by a vector of weights, a scalar (single real number), and a bias, and the neuron's transfer function. The products of the neuron's inputs and weights are summed with the neuron's bias and passed through the transfer function to get the neuron's output. The Neural net recognizes the input vectors as a set of weights on the input lines connected to a feedforward enhanced backpropagation (BP) processing unit (BP is classified as a neural net learning rule, and will be explained in a further section) that delivers the weighted outcome or target vector, ***a***.

Neurons may be simulated with or without biases. A bias is much like a weight, except that it has a constant input of 1. The constant biases are used to adjust the AVI tag travel time data (input parameters) into a form that the neural net can handle easy. The purpose of the addition bias is to reduce the relative spread of the data for each net input, n. For example if the sets of input/output range considerably, an addition bias can be added to all the output to reduce the spread. Decreasing the spread in the data reduces the training time of the neural net.

1. A scalar input (**p**) is transmitted through a connection that multiplies its strength by the scalar weight (**w**), to form the product **ω*p**, again a scalar.
2. The transfer function net input (n), again scalar, is the sum of the weighted input ***wp*** plus an optional bias (b). This sum is the argument of the transfer function *(*f).
3. A transfer function, typically a sigmoid function or a linear function, that takes the argument ***n*** and produces the scalar output, **a**.

During the feedforward process the input vector elements enter the network through the weight matrix **W**1. The corresponding bias matrix **b1** is added to the weights generating the net input, **n**.

$$\mathbf{W}=\begin{bmatrix} W_{1,1} & W_{1,2} & \bullet & W_{1,R} \\ W_{2,1} & W_{2,2} & \bullet & W_{2,R} \\ \bullet & \bullet & \bullet & \bullet \\ W_{S,1} & W_{S,2} & \bullet & W_{S,R} \end{bmatrix}, \qquad \mathbf{b}=\begin{bmatrix} b_{1,1} \\ b_{2,1} \\ \bullet \\ b_{S,1} \end{bmatrix}$$

For easy association, row indices of the elements of the matrix **W** and **b** indicate the destination neuron associated with that weight and bias, while the column indices indicate the source of the input for that weight and bias.

Figure 2.1 depicts a single layer network of ***S*** neurons with multiple input vectors, and shows how the bias effects the net input before going to the transfer function, where the transfer function is contained in the general neuron.

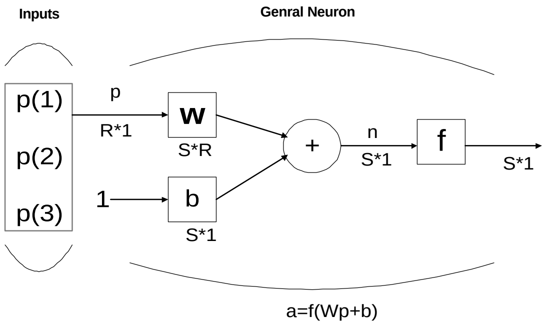

Figure 2.1 - Layer of S Neurons with R Inputs

The net input of Figure 2 can be calculated from the summed weight inputs plus a bias to form the equation:

$$n = w1,1\ p1 + w1,2\ p2 + \cdots + w1,R\ pR + b\ .$$

This expression can be written in matrix form:

$$n = Wp + \mathrm{b},$$

Where the matrix **W** for the single-layer cases can have multiple **S** neurons in that layer.

The neuron output is calculated as:

$$a = \mathrm{f}(Wp + \mathrm{b})\ .$$

If, for instance, w1,1 = 3, p1 = 2 , w1,2 = 4, p2 = 1and *b* = -1.5, then

$$a = f(\mathrm{w1,1\ p1 + w1,2\ p2 + b})$$

$$= \mathrm{f}(3(2) + 4(1) - 1.5)$$

$$a = \mathrm{f}(8.5)$$

The actual output a, influenced by the bias, depends on the transfer function. Let it be known that ***w*** and ***b*** are both adjustable scalar parameters of the neuron. Typically, after the transfer function is chosen the parameters ***w*** and ***b*** will be adjusted by some learning rule so that the neuron input/output relationship meets some specific goal.

2.5.2 - Network Architecture (Hagan et al.,1996)

Neurons that receive the same inputs and use the same transfer function may be grouped in layers. Layers of neurons may contain any number of neurons and use any transfer function. Layers may receive input from vectors presented to the network directly or from outputs of other layers. In BP, networks often have one or more layers of sigmoid neurons followed by an output layer of linear neurons. A multiple-layer neural net with nonlinear/linear transfer functions allow the network to learn nonlinear and linear relationships between input and output vectors. A two-layer network with neurons in each layer is shown in Figure 2.2

Each layer of this network has its own weight matrix, its own bias vector, a net input vector and an output vector. As shown, there are *R* inputs, *S* neurons in the first and second layer, where different layers can have different numbers of neurons. For the multiple layer networks it is easy to add the number of the layer to the names of the matrices and vectors associated with that layer. Hence, the weight matrix and output vector for layer two is denoted as **W2** and **a2** respectively.

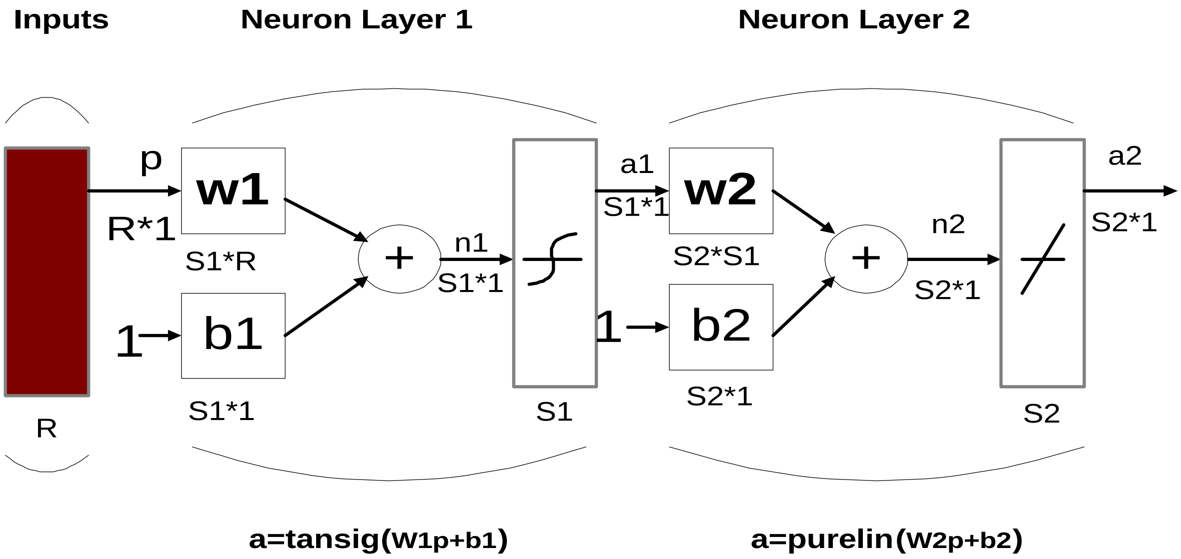

Figure 2.2 – Multilayer Tansig/Purelin Network

This network can be used for general function approximation. It has been proven that two-layer networks, with sigmoid transfer functions in the hidden layer and linear transfer functions in the output layer, can approximate virtually any function of interest to any degree of accuracy, provided a sufficient amount of hidden units are available [Hoescht, 1989]. Therefore, the neuron model key component, the transfer function, is used to design the network and established its behavior. Since a multilayer net combined with a backpropagation algorithm is more desirable for this research, the rest of the literature review will be devoted to backpropagation algorithm while the models of the multilayer net which was used in this thesis will be addressed in the *chapter 5*.

2.5.3 – Backpropagation (BP) Algorithm

The backpropagation (BP) algorithm is an extension of the *least mean square* (LMS) algorithm and was developed for training multilayer neural networks with the objective of minimizing the errors between the actual and desired output. A basic reference on this subject is “Learning Internal Representations by Error Propagation”. (Rumelhart *et al.,* 1986)

The backpropagation algorithm uses the chain rule in order to computer the derivatives of the squared error with respect to the weights and hidden layers. The learning schematics was developed by S. Grossbery in “ Competitive Learning: From Interactive Activation to Adaptive Resonance”(1997). Grossberg’s architecture indicates inputs at the first layer (inputs will be considered a layer for architectural purposes) **F1** go through the second layer **F2** and generate outputs at **F**3. While simultaneously, the expected outputs are being fed (by an external teacher) to an error signal **F**4, where the difference between the expected output and the actual output, multiplied by the derivative of the actual output, generate a different error signal **F**5. This error signal is used to change the weights in **F2- F3** pathways. These weights are then communicated to **F4-F5** pathways where they are multiplied by the **F4** error signals to generate weighted error signals at **F**5. These, appropriately weighted by derivatives as in the layers above, are then used to alter the weights in the **F1-F2** pathways. Note that the requirement that the outputs be converted to derivatives of outputs at each layer. Hence layers **F6** and **F7** adds to the complexity of the entire scheme.

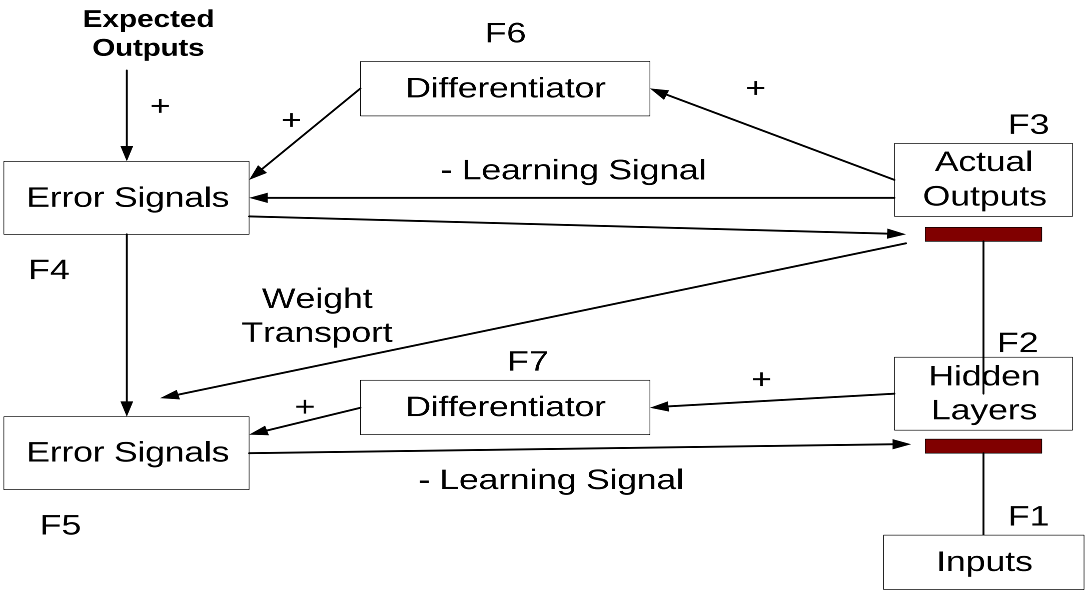

Figure 2.3 – Backpropagation Learning Schematics

(Source from “Competitive Learning: From Interactive Activation to Adaptive Resonance”, Cognitive Science, 11, 23-63(S. Grossberg, 1987))

To minimize the squared error during training, several nonlinear minimization algorithms such as the steepest-descent algorithm, Newton's method, and conjugate-gradient are generally used. In this thesis, a Levenberg-Marquardt backpropagation algorithm (LMBP) was incorporated along with Approximate Steepest Descent Rule and Gauss-Newton Method. A detailed discussion of this LMBP is available in *Chapter 5*.

2.6 – Summary of Literature Review

In recent decades, researchers have been actively investigating numerous technologies, ranging from sensitive acoustic devices to pattern recognition algorithms, in an effort to improve upon existing traffic surveillance methods and further model AVI system. In light of recent research efforts, AVI appears to show more promise as a more reliable and accurate method of predicting travel time information than other technologies, particularly loop detectors.

This work serves as a further modelling of existing AVI system using San Antonio, TX as an example. Several basic issues addressed in this literature review will be discussed in more detail, including AVI site location optimization and travel time forecasting. Detailed literature review of these topics will be performed in *chapter 4* and *chapter 5*. The Neural Network Technologies were introduced in this chapter to help readers to understand how it works. The numerous references cited in this literature review also indicate that AVI technology likely possesses untapped potential in other ATIS applications, particularly incident detection and travel time forecasting. It is hoped that this work will serve as a further step to model and improve such a system.