USING AUDIO-BASED SIGNAL PROCESSING TO PASSIVELY MONITOR ROAD TRAFFIC

by

Orla Duffner, B.Eng

Supervisors: Dr. Noel Murphy and Dr. Sean Marlow

Centre for Digital Video Processing
and School of Electronic Engineering
Dublin City University
July, 2006

DCU

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Orla Duffner

Orla Duffner

ID No.: 95669574

Date: 21/09/2006

ACKNOWLEDGEMENTS

I wish to extend my heartfelt gratitude to the many people who have enabled and supported the work behind this thesis. Special thanks goes to my supervisor, Dr. Noel Murphy, who has been an excellent and tactful mentor, dedicating many hours to this work. Similar thanks are due to Dr. Sean Marlow, Dr. Noel O'Connor and Professor Alan Smeaton for their individual advice, support and suggestions at different stages of the work. I was lucky to experience a range of unique contributions, ensuring a more rounded perspective and training.

I am indebted to Joe Marks, Paris Smaragdis and Bhiksha Raj. The timely commencement of collaboration between the CDVP and Mitsubishi Electric Research Laboratories resulted in Dr. Joe Marks extending an invitation to visit their R&D offices in Boston. His input together with advice from research scientists Drs Paris Smaragdis and Bhiksha Raj inspired the topic, direction and early stages of this work.

This work would not have been possible without the financial support provided by Science Foundation Ireland via the Adaptive Information Cluster under Grant No. 03/IN.3/I361.

On a personal note, so much thanks goes to all the members of the Centre for Digital Video Processing for technical discussions, ongoing encouragement and counselling. With such a large group it is unfair to name individuals, but I truly appreciate many different people for a large variety of reasons. Likewise for all the students and staff I met in DCU over the years, including those from the Postgrad Soc, Faculty and basketball.

Thanks so much to friends worldwide outside DCU, you have been been a lifeline for keeping in touch with non-academic life and maintaining some perspective. To my house-mates, friends and fellow PhD students Emira and Es, we've finally all finished!

So much appreciation goes to my parents and sisters; your unquestioning support

and encouragement gave me the confidence and strength to pursue and finish not only this work, but so many things in life. I will do my best to reciprocate and pass on your positive energy.

Jeroen, thank you for unconditionally believing in me. Wat ons schijnbaar hield gescheiden, ons werklijk innig te vereenen scheen.

If I've realised nothing else during this time, it's that the love of those who care surpasses all else in value.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The adaptive management of vehicular traffic on roads is a key transportation application. Sensors are required to provide information describing the behaviour of traffic in the region to be monitored. There is scope for a low-budget, efficient and robust traffic monitoring system. The hypothesis is that an audio-based approach provides a highly economical and efficient solution to monitor road traffic.

The main contributions of this thesis may be summarised as follows. In order to determine their behaviour over time, individual vehicles are successfully tracked with an efficient source localization technique based on acoustic information. The vehicle source location is determined by the inter-signal time delay of two cross-correlated microphones, known as the time delay of arrival (TDOA) localization method. A moving source model is derived from first principles to simulate the time-delay pattern due to changes in source location as a vehicle approaches and passes the array. Using the moving source model, two novel pattern extraction methods are developed to extract vehicle events and parameter values from the cross-correlation array. The first method minimizes the amount of cross-correlation data stored by extracting and tracking the path of predominant peaks, then comparing the path behaviour to the derived model to determine vehicle parameters. The second method draws on image processing techniques to search for regions or shapes of high correlation in the array that match the time-delay shape model of a passing vehicle.

Each method was tested with real traffic data of 2,267 vehicles recorded at 5 locations under a range of conditions. The shape-matching approach yielded the highest accuracy of 93% for vehicle detection with a velocity tolerance of ± 19 km/h. The positive experimental results indicate that the preferred method is a viable, economical audio-based traffic monitoring sensor system.

GLOSSARY OF VARIABLES

α	sound attenuation coefficient for atmospheric absorption
c	speed of sound
d	distance along the road from t_{ref}
D	perpendicular distance between the centre of the road and the microphone axis
δ	Dirac delta function
f_s	sampling frequency
$G_{x_1x_2}$	cross power spectral density
$G_{x_1x_1}$	auto power spectral density
k	integer constant
L	distance from the sound source to the microphone
L_w	window length
m	distance between microphones
n	additive noise
O_w	hop size
r_xx	autocorrelation function
r_{12}	cross-correlation sequence
ρ_{12}	normalized cross-correlation sequence
s	source sound signal
t	time
t_{ref}	reference time where $L_1 = L_2$, i.e. when $\tau = 0$
τ	time delay between microphones
$\Delta\tau$	rate of change of time delay
θ	angle between the sound source and the microphone
v	source velocity
ψ	weighting factor
x	audio signal received at the microphone
X	frequency-domain spectral estimate of x
X'	complex conjugate of X

GLOSSARY OF MNEMONICS

ARMA	Autoregressive moving average
AVI	Audio video interleave
CAD	Computer-aided design
CCTV	Closed-circuit television
CDB	Constant directivity beamformer
CPA	Closest point of arrival
CPU	Central processing unit
CW	Continuous wave
dB	Decibels
DFT	Discrete Fourier transform
DOA	Direction of arrival
DSP	Digital signal processing
FFT	Fast Fourier transform
FHWA	Federal Highway Administration
FMCW	Frequency-modulated continuous wave
GCC	Generalized cross-correlation
GPS	Global positioning system
Hz	Hertz
ISO	International Standards Organisation
ITS	Intelligent transportation systems
ML	Maximum-likelihood
MUSIC	Multiple signal classification
PCM	Pulse-code modulated
PHAT	Phase transform
RF	Radio frequency
SNR	Signal-to-noise ratio
TDE	Time delay estimate
TDOA	Time difference of arrival
TTI	Texas Transportation Institute
WAV	Waveform audio file format standard

Chapter 1

Introduction

1.1 General objective

While standing beside a busy road with our eyes closed, humans can easily guess what kind of traffic is passing. Is it a traffic jam, or a rapid series of vehicles at high speed? Are there many heavy trucks or motorcycles travelling in both directions, or is the road completely empty? There is a rich source of information available from any audio signal if one is capable of extracting the information, as humans are. When acoustical monitoring without human intervention is required, the audio signals can be captured via microphones and processed digitally using dedicated digital signal processing (DSP) hardware and algorithms to extract the required information. Such a system can theoretically be used to monitor acoustical environments such as busy roadsides, emulating the abilities of humans in making sense of the surrounding acoustical environment. The acoustical environment to be considered in this thesis is where road traffic noise occurs. The behaviour of individual vehicles, such as direction and velocity is the information we wish to try to extract. In summary, the general objective of this work is to passively measure audio signals with an economic microphone system to determine road vehicle traffic flow.

1.2 Contextual description of application area

Intelligent Transportation Systems (ITS) is a general term used to describe the applications of advanced technologies to air, sea and road transport. The purpose of ITS is to help monitor and manage traffic flow, improve safety, enhance mobility, save energy and promote productivity of air, sea and land transportation systems. ITS encompasses a broad range of electronic technologies, communications-based information and control systems. Such systems provide the means for professionals to collect, analyse, use and archive relevant transportation data. Improvements in traffic management, information for transport users and incident detection improve safety as well as providing users with significant reductions in total costs and travel time.

Across the globe, research and development activities are creating new products and services which industry is rapidly bringing to the market. Public authorities are increasingly embracing ITS. Growth in ITS is expected to continue, with one prediction stating that all modes of world traffic volumes will double from the 1990 level of 23 trillion passenger kilometers, to 53 trillion passenger kilometers by the year 2020. It is then expected to double again by the year 2050.

A central ITS application is the adaptive management of vehicular traffic on roads. Management tasks include traffic calming, imposing speed restraints and redirecting traffic during adverse conditions such as bad weather or heavy traffic jams. Furthermore, early incident detection may be used to alert emergency services to a crash as well as inform public road users of dangers ahead. More basic traffic monitoring systems may simply require the passive counting and classification of the type of vehicles on a road. Central to any traffic management system is the need for a constant inward flow of information describing the amount, type and and behaviour of traffic throughout the network of roads in the region to be managed. A range of different sensors exist to capture such information, based on diverse technologies. The information captured by these sensors is then transmitted to a central management system that interprets and uses the data as the basis for its decision-making process. These traffic sensors form the backbone of any traffic management system, since all successive stages are entirely reliant on the information provided by the sensors. Therefore it is of paramount importance that a traffic management system incorporates sensors suitable for its objectives.

The objectives of a traffic management system vary according to the characteristics of the road type to be managed. As the worldwide population quantity and geographical spread increases yearly, so does the length of new roads being built. The type and function of roads range from large multi-lane motorways connecting large cities in developed countries, to rural roads accessing isolated communities located in areas of extreme environmental conditions such as deserts or mountain plateaus. The objectives in managing such a diverse range of road types differ greatly and consequently dictate a range of technical and financial requirements. A spectrum of traffic monitoring approaches are required to cater for different requirements.

The project described in this thesis is based on the premise that there is scope for a low-budget, efficient traffic monitoring system that uses sensors robust enough to operate in a range of environmental conditions. Microphones and digital signal processing (DSP) chips are very inexpensive and readily available with a wide range of specifications and physical sizes. These can be used to form the basis of an audio sensor system to monitor traffic. Such a system is not expected to exceed or replace the capabilities of existing accepted technology, but rather provide an economical alternative to enhance the suite of choices available. Depending on the situation, an audio-based system can provide an optimal economical solution.

Environmental monitoring is an active research area, from people recognition and intruder sensors to object tracking. Sound source tracking systems are already successfully used in military and civilian applications to detect objects such as submarines, tanks, airplanes and people speaking in a room. Specifying the object as a road vehicle changes the application to that of traffic monitoring. The sound characteristics and behaviour of road vehicles differ from other applications. Although there has been much research on defining and classifying the characteristics of vehicular sound, there have been substantially fewer research publications oriented towards detecting vehicles. To the best of the author's knowledge, no prior research publication exists that describes a fully automatic method of audio-based traffic monitoring. This lack of prior research combined with the growth in ITS provides clear motivation for exploring the possibilities and limitations of audio-based traffic monitoring.

1.3 Challenges of audio-based traffic monitoring

There are a number of diverse challenges to be confronted when using audio data to automatically monitor vehicular traffic. The use of acoustic information to monitor traffic relies on the measured audio signal containing sufficient information to render it useful. One of the difficulties of successfully capturing the required information is contending with the physical challenges of outdoor sound propagation. These challenges include sound attenuation, distortion, masking and reflections [27, 1, 74, 119, 12]. Furthermore, the measured audio signal may consist not only of the desired information, but also noise sources and extraneous sounds. The challenge is to optimize the measurement to avoid the information of interest in the audio signal being lost amongst noise and propagation effects.

There would be little difficulty in automatically monitoring traffic if the type of traffic passing a sensor is well-separated, consisting of equally spaced homogeneous vehicles, with the same velocity and emitted sound amplitude. Uncontrolled traffic behaviour complicates the interpretation of data, which forces the system to deal with a range of event types, rather than the single event of an isolated vehicle passing the sensor. Examples include when multiple vehicles simultaneously pass the sensors in the same direction, or when successive vehicles pass in close proximity, making them effectively indistinguishable. Furthermore, sounds from an overtaking vehicle travelling at a higher velocity will be detected together with the slower vehicle being overtaken. There is a need for separating not only the information of interest from extraneous noise, but also to identify individual sound sources of interest within the measured signal. The separation and localization of sound sources is an active research area and ongoing challenge.

A traffic management system uses input data such as the amount, direction and velocity of individual vehicles. Therefore, any sensor system providing data for the purpose of traffic monitoring or management should be designed to provide such data. Simply detecting a relevant sound source is insufficient, and it is necessary to track the sound for a limited time to establish its behaviour and hence movement characteristics. This research area, called source tracking, is a natural extension of source localization. To identify the movement characteristics and hence the vehicle parameters from tracking a sound source, the patterns created by the source need to be analyzed. In the context of automatic vehicle monitoring, the application of pattern recognition is based on an understanding of the underlying factors causing the observed patterns and how these factors translate to the parameters of interest. These steps can be redescribed as two tasks; (a) model the relationship between variable parameters and a moving source, (b) apply the model to measured moving source patterns, to determine the source parameters. Both tasks form the challenges of pattern recognition of a moving sound source in vehicle tracking. In this manner, the parameters of vehicles passing the sensors may be determined and transmitted to traffic management systems. The challenges of using audio data to automatically monitor vehicular traffic aid in defining the objectives of this thesis.

1.4 Research goals

The overall objective of developing an economical audio-based road traffic monitoring system is now expressed as a series of defined research goals.

1. Investigate the viability of using audio information to monitor traffic by testing a number of different signal processing approaches;
2. Develop an audio-based traffic monitoring system that uses a small, economical and compact microphone array system (preferably just a pair of separated microphones) and simple signal processing algorithms;
3. Derive a series of mathematical equations that model the data pattern generated by a moving vehicle. Use this model to simulate a moving source for a range of system parameters, thereby determining the optimum parameter values;
4. Passively detect, separate and track multiple vehicles solely based on measured audio data;
5. Automatically determine the behavioural parameters of a vehicle from the pattern of data measured and extract the relevant traffic characteristics, to include quantity, direction and velocity of vehicles.

1.5 Main research contributions

The research carried out during the course of this thesis presents a number of significant contributions to the area of passive traffic monitoring and audio event detection. Three novel traffic monitoring systems of varying accuracy were developed and evaluated. Although a traffic monitoring approach based on cross-correlation data has been described in publications, no fully automatic system is presented. Prior experiments are often performed based on simulations as opposed to actual traffic data. Furthermore, none of the previously published systems are fully automatic, since they rely on manual visual analysis of time-delay patterns to detect a passing vehicle. Novel, fully automatic vehicular pattern recognition systems were developed and are described in this thesis. The pattern extraction was found to be a central research challenge, which may contribute to future time-delay pattern analysis research.

Early experiments were performed to test the use of a large range of existing audio features in detecting the presence of a vehicle. These features included the average zero-crossing rate, signal energy, spectral centroid and fundamental frequency. Audio features are typically used in sound classification and separation, where there is little noise and the temporal-spectral characteristics of the sound being examined are distinctive and distinguishable. They are generally not designed to be robust to uncontrolled outdoor environments or to detect sounds whose characteristics are often negligibly different to the background noise. It was found that sound amplitude was the only feature vector to change noticeably in the presence of a passing vehicle. It was concluded that the examined audio features are not suitable for traffic monitoring, so an audio feature-based approach was not considered further. A description of investigations involving audio features is included in Appendix A to justify this decision.

Each traffic monitoring system developed in this project and described in this thesis, draws at some point on prior techniques. However, the techniques are implemented and combined in a novel manner that maximizes system performance to attain the research objectives. For example, one system utilizes a weighted cross-correlation array to determine time-delay values. A shape-matching pattern extraction approach is then applied to detect the presence and parameters of models in the cross-correlation array, where the presence of model shapes indicate a passing vehicle. The most suitable algorithm for obtaining the cross-correlation array for passing vehicles is based on prior methods. Similarly, a known shape-matching pattern extraction technique used in the field of image processing was found to be the most robust, accurate approach to detecting vehicles. However, a variation of the shape matching technique concept was developed and applied in a novel manner to the most optimal cross-correlation representation of vehicular traffic. A major research contribution of this work is therefore based on the manner in which the most suitable methods for audio-based vehicle tracking were combined, implemented and evaluated.

A series of mathematical equations were derived by the author to model the passage of a vehicular sound source on a road as measured by a pair of microphones. Variables defining environmental and system parameters that affect the shape of the moving source trajectory were included. In this manner, the model was used to perform simulations of system performance with a range of values. The derived model and simulation results may be used in future research to rapidly choose an appropriate system geometry and parameter values without the need for systematically testing the effect of each parameter change.

Audio data from a substantial number of different types of vehicles was recorded at a range and quantity of different locations. A manual list of each vehicle event was created for up to three different types of reference data. This large body of reference data based on real traffic was used to evaluate the developed systems in a series of original experiments in unique conditions. The experimental results contribute to a more in-depth measure of audio-based traffic monitoring system capabilities, under the circumstances of the experiments performed.

1.6 Thesis structure

The thesis is broadly laid out in three general sections. The first section contain relevant background information on traffic sensors, sound propagation and source localization in Chapters 2 to 4. The second section details the research approaches taken, and how they were implemented during as described in 5 to 7. Chapters 8 and 9 form the final section, describing the experiments performed, outlining the results and conclusions with a critical review of the thesis and making recommendations for future research. A summary of each chapter is now provided.

Chapter 2 provides a comparative discussion of existing traffic sensors and presents results from the literature of a series of extensive experiments, many of which were performed by US Government-funded Transportation Organisations. Chapter 3 starts with a discussion of the acoustical effects of outdoor sound propagation and possible implications this may have on audio traffic monitoring. It then moves on to describe the characteristic noise of different road traffic and how this varies under different conditions. Chapter 4 describes different sound source localization techniques and their individual merits, before justifying the approach used.

Chapter 5 details the implementation of a robust weighted time-delay of arrival (TDOA) cross-correlation approach to determine vehicle source location relative to a microphone array. Individual vehicles occupying separate locations can be distinguished in the generated cross-correlation array. Chapter 6 provides a mathematical framework describing a moving sound source and derives appropriate equations that can be used to simulate and model such an event. The equations are then used to model different scenarios and demonstrate the implications of the choice in parameters on system performance and accuracy. Three automatic pattern extraction techniques to analyse the amplitude or cross-correlation data from Chapter 5 are described in Chapter 7. Section 7.1 describes a simple loudness-based traffic monitoring approach. Although limited in use, it is considered useful to consider such a method in order to compare results against more sophisticated approaches. Each of the three pattern extraction techniques determine the parameters of passing vehicles.

Chapter 8 describes the experiments conducted to evaluate audio traffic monitoring systems. A large set of reference data based on real traffic is used to compare the developed vehicle tracking methods. The results are critically analysed to determine the performance of each system. Finally, Chapter 9 concludes the thesis with a summary of observations, discussion of lessons learned and recommended future work.

Chapter 2

Traffic Sensors in ITS

In order to evaluate audio traffic sensors in context, it is necessary to be aware of the capabilities of existing traffic sensing technology. Some sensors will outperform others in certain conditions, but at costs that may be unacceptable for certain applications. To effectively compare different technologies they should be exhaustively tested over a long duration in the same operating conditions. There is a broad range of publications documenting state-of-the-art traffic sensors with comparative evaluations. Relevant literature is described in this chapter to provide a critical analysis of existing technology. A comprehensive selection of existing traffic sensor technologies are described in this chapter. Section 2.1 describes types of existing traffic sensors and the fundamental technology involved. Extensive evaluations and comparisons of different traffic sensors have been carried out, many of which are summarized in Section 2.2. Conclusions about traffic sensors are made in Section 2.4.

Traffic sensors are strategically placed along or under the road as graphically illustrated in Figure 2.1, and communicate relevant data to a traffic management system. Such a system can only be as accurate and reliable as the data provided. Traffic sensors have been developed from a rich array of technologies such as video, radar, magnetics and acoustics. This thesis concentrates on the use of acoustic information for passively monitoring vehicular traffic.

2.1 Traffic sensors

The first known vehicle detection device appeared in Baltimore in 1928. Drivers on a side street would sound their horn to activate the device, which consisted of

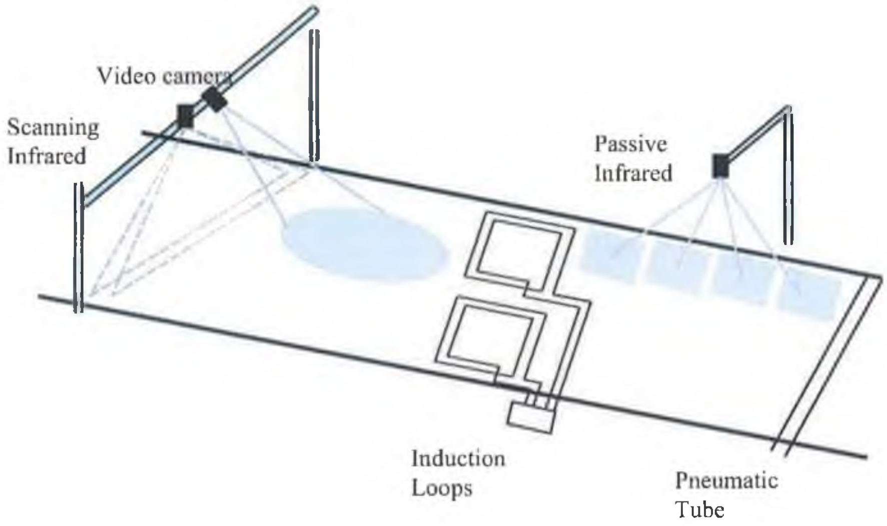


Figure 2.1: Illustration of traffic sensors along a road

a microphone mounted in a small box on a nearby utility pole. Another device introduced around the same time was a pressure-sensitive pavement detector using two metal plates acting as electrical contacts forced together by the weight of a passing vehicle. Proving more popular, it enjoyed widespread use for over 30 years [138]. Inductive loops were introduced in the early 1960s and have become the most widespread detection system to date. However, problems with inductive loops and progress in technology, have led to the introduction of numerous non-intrusive devices which utilize a variety of technologies to address the failures of inductive loops. The devices can be categorized as intrusive or non-intrusive, passive or active.

Intrusive sensors such as induction loops, passive magnetic sensors and pneumatic tubes must be placed on or under the road [91]. As a result, it is necessary to temporarily close the lane for installation and maintenance. Multiple intrusive sensors are required to monitor multi-lane roads. On the other hand, non-intrusive sensors are typically placed adjacent to or above the road of interest and in some cases a single sensor can monitor multiple lanes. Non-intrusive traffic systems largely consist of three parts; a sensor to electronically capture relevant data, a microprocessor to digitize and process the data and software to interpret the raw information and convert it into traffic information suitable for communication to traffic management systems. Appropriate sensor placement and elevation is critical to the system performance. Sections 2.1.1 to 2.1.9 presents some popular technologies and research activities for vehicle detection.

2.1.1 Induction loops

The induction loop detector is the most widespread and mature traffic sensing technology. It is embedded under the road surface in the middle of a traffic lane as illustrated in Figure 2.2. An induction loop consists of a loop of insulated wire connected to an oscillator circuit. The wire loop is excited with an AC signal ranging in frequency from 10kHz to 200kHz, and functions as an inductive element [88]. When a ferrous vehicle passes overhead, the circuit inductance changes, due to eddy currents induced in a metal vehicle. The decreased inductance causes an increase in the oscillation frequency, which prompts the electronics unit to send a pulse to the controller. Induction loops require a small current to operate, causing them to be classified as active magnetic devices. The loop shape and size depends on the detection purpose at that location. The higher the number of windings in the loop, the greater the loop sensitivity in detecting ferrous objects.

Induction loops can detect the presence and passage of a vehicle to provide accurate data on the number of vehicles and lane occupancy for most historical traffic management applications. A single loop detector cannot directly measure speed or density. For this reason, two separate loops are often used where the differential detection time and known inter-loop distance can be used to determine vehicle speed. Vehicle classification can be performed by estimating the vehicle length. Such systems are used at toll plazas to determine the payment due based on vehicle size. Reliability is a major issue with induction loops even though there have been improvements through better packaging and installation techniques. The high failure rate is due to a combination of factors; poor installation, poor materials and road deterioration. Due to their intrusive nature, induction loops require lane closure for installation

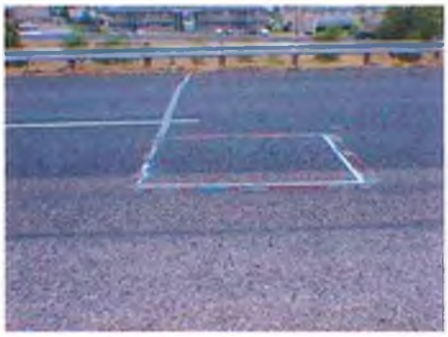


Figure 2.2: Induction loop embedded in a road

and maintenance, which can be an issue in certain locations.

There is widespread use of loop detector systems in Europe for traffic detection and monitoring. The Dutch report an extremely high reliability rate for inductance loops, perhaps because they developed their own specifications after determining that commercially available systems did not meet requirements for reliability and long-term operation [127]. Gajda et al [63] described using induction loops to classify vehicles using their magnetic profile as well as to detect the number of axles and to measure distance between them.

2.1.2 Passive magnetic sensors

Passive magnetic sensors detect the disruption in the earth's natural magnetic field caused by the movement of a vehicle through the detection area. To detect this change, the device must be close to the vehicle and is usually installed under the pavement. There are two types of magnetic field sensors; fluxgate magnetometer and the induction or search coil magnetometer [91]. Both types of magnetic sensors are intrusive and require the road to be cut or tunneled. However, they are less susceptible to the stresses of traffic than loops.

A magnetometer was introduced in the 1960s as an alternative to the inductive loop detector in specific situations. The vertical fluxgate magnetometer detects changes in the vertical components of the earth's magnetic field, while the two-axis fluxgate magnetometer detects changes in the vertical and horizontal components of the earth's magnetic field. It generally consists of primary and secondary windings surrounding a high-permeability soft magnetic core. The secondary windings are offset by 90^o to sense the horizontal and vertical magnetic fields and are usually aligned with the direction of traffic flow. The output voltage increases when a vehicle is in the detection zone. When operating in the 'pulse output' mode, the passage of a vehicle can be measured, while in the "presence" mode a continuous output is given as long as the voltage exceeds a threshold. For a vertical axis magnetometer to function, the vertical components of the earth's magnetic field must exceed 0.2 oersteds, therefore vertical axis magnetometers cannot be used near the equator where the magnetic field lines are horizontal. It is possible to separately detect two vehicles a foot apart, making the magnetometer more precise than the induction loop detector for counting vehicles.

Search coil magnetometers (or magnetic detectors) consist of a highly permeable magnetic core, on which are located several coils in series, each consisting of a large number of turns of fine wire. A voltage is induced due to changes in the magnetic flux lines with respect to time. The coil axis is perpendicular to traffic flow and disturbed magnetic flux lines cut the turns of the coil as long as a vehicle is in motion through the zone of influence. As a result, such units do not function as presence detectors, requiring some minimum vehicle speed for detection (e.g. 5 to 16 km/h). Models vary in size and unlike induction loops can be fastened to the underside of a bridge where steel is present, embedded in the road or flush-mounted with the road surface. Using a magnetometer requires far less road cutting and they tend to survive longer than induction loops in brittle pavements. Recently Nishibe et al. [132] proposed on-road lane markers with a built-in magneto impedance sensor and power source. One advantage of such a system is that they would not need to be embedded in the road.

2.1.3 Pneumatic tube

Pneumatic sensors consist of tubes of rubber filled with compressed air that are placed across the surface of the road perpendicular to traffic flow. The impact of a vehicle tyre causes a burst of air pressure along the tube, closing an air switch and hence producing an electrical signal that is transmitted to a counter. It is a portable device, usually used for short-term traffic analysis and research since it eventually wears out. By counting the quantity and distance between axles, vehicle classification can be performed. With more than one pneumatic tube, the vehicle velocity can be indirectly estimated. They are quick to install, economical and simple to maintain.

2.1.4 Traffic sensor using piezoelectric material

A piezoelectric material is a specially processed material capable of converting kinetic energy to electrical energy. Some polymer materials exhibit these properties. The piezo-electric traffic sensor is coaxial with a metal, braided core element, followed by the piezo-electric material and a metal outer layer. It is subjected to an intense electrical field during the manufacturing process, which radially polarizes the material. It changes the amorphous polymer into a semi-crystalline form, while retaining many of the flexible properties of the original polymer [68]. When a vehicle passes over the sensor, the mechanical impact or vibration generates electrical charges of opposite polarity at the parallel faces, inducing a voltage. The measured voltage is proportional to the force or weight of the vehicle and decays if the force remains constant. Piezoelectric sensors can be used to classify vehicles by axle count and spacing, and to measure vehicle weight and indirect speed as part of weigh-in-motion systems.

2.1.5 Video imaging systems

Originally video cameras required a human operator to interpret closed-circuit television (CCTV) images. Technology has progressed to current video applications that automatically analyse and interpret the video based on image processing techniques. Various video traffic sensing systems are currently produced, used and tested in every-day situations. These are compared against other traffic sensors in Section 2.2. The development of algorithms to analyse video data for traffic monitoring continues to be a large research area, with a range of potential applications. Video sensors could be used for vehicle detection, counting, localization, tracking, recognition and classification along a motorway, in built-up areas or at an intersection. Vehicle licence-plate recognition, incident detection (such as a collision), vehicle lane-change detection, queue detection and vehicle re-identification for the purpose of journey tracking/travel time estimation are other potential applications.

There are three types of video system: tripline, closed-loop tracking and data association tracking [196, 148, 117, 108, 201, 94, 59]. *Tripline* systems monitor a limited number of user-defined detection zones. Pixel changes identify the crossing of a vehicle through a zone. *Closed loop systems* first detect, then continuously track vehicles within the camera field of view [117]. Lane-to-lane vehicle movement can thus be determined, which can be transmitted to alert drivers to erratic behaviour. *Data association tracking systems* uniquely identify areas of a particular vehicle or group of vehicles and track them from frame to frame as they pass the camera field of view [102, 199, 35]. This has the potential to link travel-time and origin-destination pair information by coordinating data from a series of cameras.

The type of camera used determines the quality and resolution of image obtained. Some cameras have automatic iris and gain controls, which adjust the light levels entering the camera and adjust the sensitivity of the camera respectively. Although required when background lighting changes, this is a disadvantage as it also responds

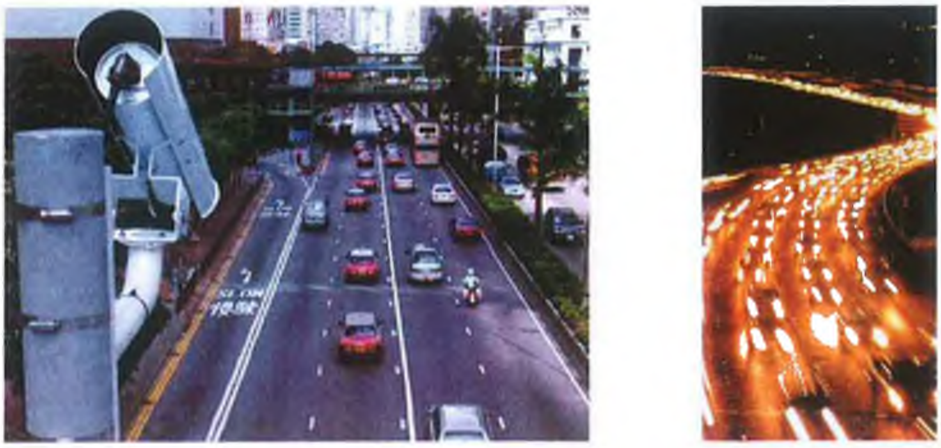


Figure 2.3: (a) Autoscope Solo Pro automatic incident detection in Hong Kong (b) Night image of vehicles to illustrate headlight blooming

to headlights, reflections and bright objects as well as objects temporarily dominating the camera's field of vision. Video cameras can be deployed to view upstream or downstream traffic. Upstream viewing can be blocked by tall vehicles, headlights may cause image blooming at night (shown in Figure 2.3) but also incidents are not blocked by resultant traffic queues. By viewing downstream, the camera can be hidden from the driver and vehicle identification is made easier at night. The measured vehicle speed accuracy depends on camera elevation, since the measurement error is proportional to the vehicle height divided by the camera mounting height. The ability of the system to distinguish between two closely spaced vehicles is also dependent on the camera mounting height. With a mounting height of 6-9m, the camera should be placed centrally over the middle of the road, whereas with a height of 15m or greater, cameras can be mounted on the side of the road. Camera motion due to high winds can be an issue with video systems. In optimal circumstances, current CCTV technology should allow viewing of 0.4 to 0.8km in each direction [91].

The software algorithms required to analyse the images can vary greatly in complexity and accuracy. There are two general approaches; model and non-model based. Non-model based systems have no knowledge of the appearance of a vehicle, simply detecting and tracking objects in the scene, while model-based systems strive to gain an understanding of the image. A classical approach to vehicle detection and tracking in video involves the subtraction of background information to create a difference image [40]. The remaining image can be analysed using techniques such as motion estimation, colour similarity and horizontal symmetry to detect vehicles and track their location over successive video frames. Karmann [87] and Zhang [200] model the background as a slow time-varying image sequence to adapt to changes in lighting and weather conditions. Since vehicles can be partially occluded in congested traffic, vehicle sub-features instead of entire vehicles can be tracked. In this manner, vehicle edges, corners and two dimensional patterns can be tracked, giving some immunity to shadows. However, performance is still deteriorated by continued full occlusion and a less-than ideal camera mounting position. Gupte et al. [67] grouped related foreground regions together to form vehicles which are classified and localized.

Some background image generation approaches fail when applied in urban traffic situations because they have some different features to highway traffic. Examples include being separated into successive blocks by intersections, and traffic conditions varying from block to block. When the traffic travels at a lower speed or even remains stationary, the background model is corrupted with noise due to stationary vehicles. Occlusion, noise, complex lighting and changes in lighting and weather conditions can cause temporary difficulties in differentiating between background and foreground. Other problems with background subtraction are that a complex background learning model is time-consuming while a simple differencing technique cannot guarantee good segmentation performance.

Non-model based systems have no information regarding the appearance of a vehicle, working by simply detecting and tracking objects in the scene. Non-model based traffic monitoring systems rely on motion detection to segment moving regions from the image, generally via frame differencing or feature-based tracking [24, 157]. Model-based systems strive to gain an understanding of the image. Models are used to represent knowledge of the appearance of vehicles and possibly the geometry of the traffic scene, usually taking the form of 3D wireframe models. Image data is mapped to corresponding 3-D model descriptions and compared, as described by Lou et al. [114]. Model-based object recognition is then used to locate vehicles in images and track them from frame to frame. Some model-based recognition methods use background subtraction [64, 201] while others analyse the entire image or regions thereof [174].

Vehicle licence plate detection is another active research area. Racal Research Ltd. describes using ordinary CCTV cameras [183]. Yanamura et al. [197] describes a method to extract and track a vehicle license plate using the Hough Transform and Voted Block matching, making it more robust to illumination changes and occlusion.

Castello [34] describes a number plate recognition method that first screens images to select those showing motion. Next the number plate is located and character segmentation, optical character recognition and context verification are performed. Finally character data from different images are fused to extract a single number plate for a given vehicle.

Some of the technical applications and research areas of video sensors in traffic monitoring are described in this section, demonstrating that the use of video sensors for traffic monitoring is a viable option. There are a range of problems that affect the usability and accuracy of video sensors, from lighting variations and occlusion to weather conditions and noise. The computational requirements for complex image analysis algorithms are significant, as is the financial cost of such a system. Section 2.2 describes objective comparative tests performed to evaluate the different traffic sensors, during which video sensors are placed in context with other options. A more critical comparison of all sensors is reserved until the end of that section.

2.1.6 Infrared

An infrared traffic system is similar to a video system in that it consists of an infrared camera, microprocessor and image processing software. It can be mounted overhead or at the side of the road. The captured energy is focussed onto an infrared-sensitive material at the focal plane, and can be in the near infrared (0.87 to 1.5μm), mid-infrared (3 to 5μm) or long wavelength band (8 to $\geq$12μm) [90]. As the wavelength increases through the infra-red spectrum, the dominant energy shifts from reflected to emitted energy. There are two types of infrared sensors; active and passive.

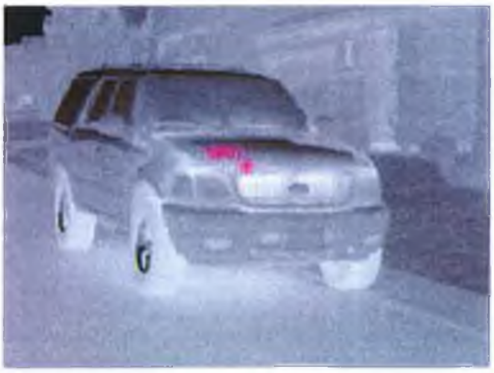


Figure 2.4: Infrared image of a vehicle Source: www.infrared1.com/gallery/

Passive infrared sensors transmit no energy, simply detecting reflected and emitted infrared energy from objects in their field of vision and the atmosphere. When a vehicle passes, the vehicle and road surface energy can be compared [78]. As a result, atmospheric temperature and weather conditions will affect the signal, particularly at the shorter infrared wavelengths. Active infrared sensors emit laser beam(s) at the road surface and measure the time for the reflected signal to return [141]. The return time is reduced when the presence of a vehicle causes reflections or scattering. Infrared sensors can measure the amount of traffic and speed as well as detect pedestrians and classify vehicles.

2.1.7 Radar

Radar is a system developed before and during World War II that uses radio waves to detect objects. The term RADAR is an acronym for *Radio Detection And Ranging.* Most roadside radar sensors operate at 10.525GHz and are limited by Government regulations to certain frequency intervals and transmission power. The radar sensor may be forward-looking with a narrow beamwidth or side-mounted with multiple detection zones, depending on the application and required accuracy. Radar devices calculate the distance to a vehicle by determining the time delay between the emitted and reflected signal. There are two types of radar used in traffic management; continuous wave (CW) Doppler radar and frequency modulated continuous wave (FMCW) radar [146, 198].

For CW Doppler radar, a pure continuous signal of a known frequency is transmitted by one antenna of the device. A second antenna receives the signal reflected from an object. There is a difference in frequency of the transmitted and received signal due to the Doppler effect. In this manner, a relative decrease in received signal frequency is due to a vehicle moving away, while a signal frequency increase is from an approaching vehicle. Only moving vehicles traveling at speeds greater than 4.8 to 8 km/h can be detected by the CW Doppler radar, where vehicle velocity is proportional to the frequency shift [91].

The FMCW radar transmits a pulsed microwave signal with constantly changing frequency in a fixed fan-shaped beam, equivalent to a long elliptical footprint on the road surface. Any non-background targets will reflect the signal back where it is compared to the transmitted signal. Vehicle presence can be directly measured for a stopped or moving vehicle. By dividing the field of view into range bins, vehicle velocity can be calculated from the time difference between a vehicle arriving at the leading edges of two bins. The Doppler principle can also be used to calculate vehicle speed, as shall be further described in Section 3.2.6.

2.1.8 Ultrasonic

Ultrasonic sensors use sound energy to transmit and detect pulses at 25-50kHz, above the human audible range. The presence of a vehicle changes the reflected signal. Constant frequency ultrasonic sensors detect the passage and velocity of a vehicle by the proportional Doppler shift in received signal frequency. Sensors operating by this principle are used in the Japanese highway infrastructure, mounted overhead and facing approaching traffic at a 45^o angle [123]. Range-measuring ultrasonic sensors measure the time delay between transmitting a series of pulses and receiving them. Pulses typically range from 0.02 to 2.5ms in width, with a repetition period of 33 to 170ms. Range-measuring sensors are more widely used than the constant-frequency type.

2.1.9 Acoustic traffic sensors

Passive acoustic array sensors detect vehicle sounds using an array of microphones aimed at the road. When a vehicle passes, the increase in acoustical energy is detected. The location of a sound source, or sources, can be determined by using a microphone array and source localization techniques. By tracking the source location over time, vehicle velocity is calculated. Vehicles can be classified based on differences in acoustical characteristics. There are two audio-based traffic monitoring products currently available for basic traffic monitoring; SmartSonic by IRD Inc. and SAS-1 by SmarTek Systems. Figures 2.5(a) and 2.5(b) present images of both audio-based traffic monitoring systems. A two-dimensional array of microphones and beamforming localization approach is used by the SmartSonic and SAS-1 systems.

The SmartSonic device measures the time delay of arrival of sound between microphones in the array. The detection zone depends on the aperture size, frequency band and array geometry. The SmartSonic is tuned to 9kHz with a 2kHz bandwidth, with a detection range of 6 to 11m. The SAS-1 traffic monitoring system is an implementation of US Patent Number 5,798,983 [99]. It forms multiple detection zones with a microphone array and signal processing, to monitor up to 7 lanes when over the road or 5 at the roadside. Every 8ms the detection zones are checked and can be adjusted to 1.8m or 3.6m at a mounting height of 6-12m with the frequency range of 8-15kHz being processed. The technology behind the system is described in Section 4.4.1. The SAS and SmartSonic traffic sensors were compared against other traffic monitoring technologies in a range of experiments, details of which are in Section 2.2.

Research on road vehicular noise and the use of sound to monitor traffic is fairly limited, especially when compared to the large quantity of publications on video-based traffic analysis. Active topics have been largely focussed on modelling, classifying and tracking vehicular noise. A brief overview of relevant literature is presented next.

Acoustic traffic monitoring research

Early research on acoustical traffic monitoring involved measuring and modelling noise generated by road vehicles, in order to examine the temporal and frequency-domain nature and levels of noise [195].

Various mathematical models for predicting road traffic noise were developed. For the first models developed during the 1960s, vehicles were assumed to be radiating the same sound power and moving at the same constant speed with equal spacings between them [84]. Calculations commonly assumed free field conditions with no Doppler effect. Later, more sophisticated and complicated statistical methods were used to introduce more realistic situations that incorporated effects such as ground

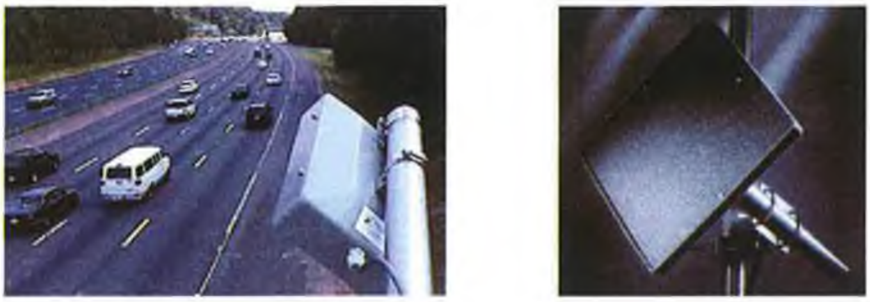


Figure 2.5: (a) SAS-1 acoustic array sensor by SmarTek Systems, Woodbridge, Virginia (b) SmartSonic acoustic sensor by IRD Inc.

absorption, the Doppler effect, atmospheric absorption and directional properties of the sound [100, 101, 181, 27, 140, 28, 60, 139]. The complicated calculations involved and limited practical interest led to a decline in the effort to produce mathematical models that could be used to define the statistical parameter of the noise along the side of a road [103]. A recent report by Jonasson [86] describes road vehicle noise measurements for use in prediction. The noise measurements described in this report are more representative of modern vehicle noise. Ban [17] et al analyses, and then synthesizes, car noise as a combination of harmonically related engine tones and broadband friction noise.

During the '70s and '80s, particularly active audio-based traffic research involved the prediction and modelling of vehicular noise. More recently the focus shifted to investigating approaches to interpret the information available from vehicular noise. Applications included vehicle detection, classification and velocity estimation. Couvreur and Bresler [45] attempted to use the Doppler effect to estimate vehicle speed and position using a single sensor. However, results were poor, due in part to background and wind noise, and also because the generated Doppler model did not account properly for all the sound wave propagation effects. Modelling vehicle acoustic signatures is a difficult problem, which can be sidestepped by including a second sensor, as demonstrated by Pérez-Gonzáles and López-Valcarce. They published a series of papers describing an approach to vehicle velocity estimation using the time delay between a pair of microphones that made no assumptions on the acoustic signal emitted by a vehicle [150, 113, 112, 111].

Vehicle recognition and classification is another area that attracts a variety of approaches. Nooralahiyan et al. [134, 135] described an approach to classifying vehicles into four broad categories using a directional microphone and linear predictive coefficients. Huadong et al. [76] characterized noise patterns using frequency vector principal component analysis to recognise whether a new sound is from a vehicle of known type for subsequent classification. Recording was found to require stable conditions and high performance equipment to build a reliable signature library. Since vehicle-generated noise is constantly changing under different conditions as technology progresses, such a sound library would quickly become out of date. Other avenues of vehicle recognition research include discriminating between aircraft and land vehicles [149].

Vehicle detection and tracking is popular both for traffic monitoring applications and military situations. The use of wideband array processing algorithms for acoustic tracking and classification of ground vehicles, such as army tanks, is described by Pham et al [152, 151]. Although the approach is relevant, the characteristic sound of such ground vehicles is significantly different to that encountered on typical civilian motorways.

Forren and Jaarsma [62] describe a tyre-noise based traffic monitoring approach for urban roads. It uses a microphone array to localize the sound source by means of cross-correlation where Doppler compensation is included. It was necessary to manually locate the vehicle correlograms in the data as the entire process was not automated. The vehicle location and velocity could then be obtained as well as vehicle type, based on length and number of axles. An array-based traffic monitoring technique applied to urban situations was described by Chen et al. [38, 39] which uses a cross-correlation based algorithm. Similar to Forren, Chen did not extract the traffic indicators automatically from the data but relied on manual intervention. Nevertheless, the cross-correlation approach described by Forren and later Chen is closely aligned to work described in this thesis, and will therefore be described and compared in detail in Section 4.4.2

2.2 Traffic sensors evaluation

Extensive field tests have been performed by a number of different organisations to compare different traffic sensors. Numerous research and government publications are available, providing an exhaustive study of relevant technology. Four of the most relevant field tests and their findings are described in this section, with observations on the results being made in Section 2.2.5.

2.2.1 Hughes Aircraft Company

Initiated by the U.S Federal Highway Administration (FHWA), Hughes Aircraft Company conducted a large-scale evaluation of non-intrusive technologies between 1992 and 1995 entitled *Detection Technology for IVHS*[92]. In a variety of weather conditions over 27 different sensors were deployed, including video, CW Doppler radar, FMCW radar, laser radar, passive infrared, ultrasonic, passive acoustic, magnetometer, magnetic and inductive loops.

The sensors were evaluated in terms of performance only - cost was not taken into account. Induction loops were found to be the most consistently accurate detectors for vehicle counting while video, magnetometer and microwave detectors showed a great deal of promise. Target accuracies were specified that support future ITS applications, in order to benchmark the evaluated sensors. A target velocity measurement accuracy of ±1.6km/h was found to be beyond the ability of most detectors.

Doppler microwave detectors were able to support the 8km/h speed accuracy requirement on a per vehicle basis, but were unable to detect stopped or slow traffic. For slow-moving traffic, video and microwave or laser radars may be required. It was explicitly stated in the report that each technology has strengths and weaknesses imposed by physics that governs its operation, causing a specific technology to be wholly unsuitable or ideal for a particular application. Subsequently, it was claimed that there is no "best detector".

2.2.2 Minnesota Department of Transportation evaluation

Between 1995 and 1997 the Minnesota Department of Transportation conducted a two-phase evaluation of 25 sensors consisting of eight technologies (magnetic, sonic, ultrasonic, microwave, radar, infrared and video) for the FHWA. The purpose was to analyse device capabilities and performance (as opposed to device-by-device comparison), in a wide variety of weather and traffic conditions including rain, sleet, snow and high winds [96]. The Smartsonic acoustic sensor was tested in a position adjacent to and above the road.

The importance of considering more than just performance and cost when comparing traffic sensors was described in the conclusions of the above report. Relevant factors to consider included intended use, ease and flexibility of installation, mounting location, communication capability, power requirements, available traffic information and the impact of weather on performance. Video devices were found to require extensive installation and calibration work before use. The video and passive acoustic devices counted vehicles with an error margin between 4 and 10% of baseline traffic volume data. Pulse ultrasonic, doppler microwave, radar, passive magnetic, passive infrared and active infrared were found to count vehicles with an error margin of 3% or less. All the device speed measurements demonstrated a maximum error margin of 8% of the baseline speed data, with radar, doppler microwave and video being the most accurate.

2.2.3 California Polytechnic State University sensor evaluation

In 1999 the California Polytechnic State University assessed advanced imaging technologies for potential application to roadway surveillance and detection, particularly using wavelengths longer than the visible spectrum in adverse conditions of fog or dust [116]. Ten types of infrared, one millimeter-wave still-frame and an visible image video system were used during experiments. For scenes without fog, the visible camera performed best, followed by the 3-5 μm cameras. Under conditions of light advective or radiative fog, the 3-5μm camera performed best, with the visible still giving a strong relative performance. Under all the conditions that did not include heavy fog, the 8-12 μm cameras evaluated provided the poorest vehicle detection. It was concluded that there are a limited number of situations for which non-visible spectrum imaging is justified. Infrared and millimeter-wave imaging technologies of the time provided marginal or no net advantage compared with conventional colour CCD video cameras for typical surveillance needs. As these technologies mature and improve and costs decrease, they may prove more attractive either as stand-alone systems or in a sensor fusion capacity.

2.2.4 Texas Transportation Institute sensor evaluation

Between 1998 and 2002 two consecutive evaluations of vehicle detection systems were performed by the Texas Transportation Institute to examine the performance, characteristics, reliability and cost of different technologies. Video, radar, acoustic, magnetometers and inductive loops were considered. Some of the equipment used is shown in Figure 2.6.

The first research project took place from October 1998 to February 2000 and is reported in [126]. Ease of setup and calibration, cost and parameter accuracy were the three evaluation criteria. The video system was by far the most difficult to set up and calibrate. During loss of power, the video system required being physically reset for it to resume operation. The installation cost of the acoustic system was significantly less than for video or magnetometer systems, and was found to be economically

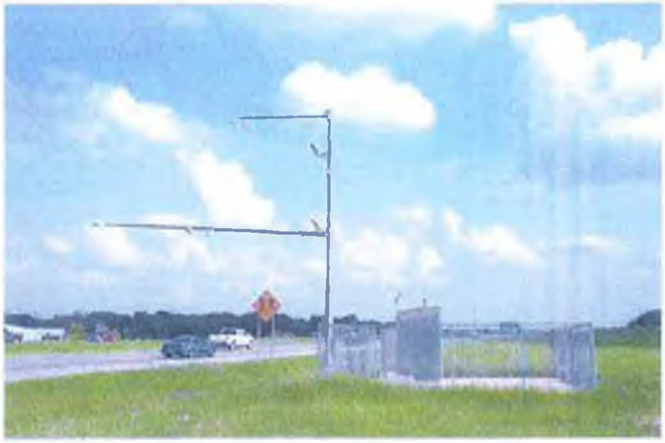


Figure 2.6: TTI Freeway detector test bed, Image Source: [126]

attractive on a per-lane basis, since it can monitor up to five lanes. The acoustic system only cost $4000 while the magnetometer system cost $9900 and the video system cost $13,200. In terms of parameter accuracy, the magnetometer was the only one of the three detectors unaffected by rain; it also demonstrated the best speed accuracy of the three systems. For 95% of the time, the magnetometer and acoustic system predict velocity within an error margin of 8mph and 11mph respectively. The video and acoustic systems demonstrated significantly worse speed performance during wet weather, with the measured acoustic speeds spuriously increasing by 10mph compared to dry weather measurements. The video system performance at night was unacceptable, partially due to a lack of street lighting.

A second TTI evaluation of vehicle detectors took place between February 1999 and August 2002. An inductive loop system was used, as well as radar, acoustic and video detectors. The non-intrusive devices were compared based on speed, count and occupancy parameters. As reported by Middleton and Parker, [127], relevant findings are described as follows. The inductive loop system classification accuracy based on a data-set of 1,923 vehicles was 98.9%, with an almost perfect count accuracy. The video system provided the most consistent performance of the non-intrusive traffic sensors, however it was the most expensive. Count accuracy was within 10% until speeds dropped below 40mph, when the error increased to between 10 and 25%. Speed estimation was excellent with an error of between 0 to 5mph. The most accurate occupancy information obtained from the video sensor was a difference of less than 1%. The radar system tested employed the FMCW principle described in Section 2.1.7 and had lowest life-cycle cost for freeway applications [125]. The count accuracy was always within 10% and the speed accuracy was excellent except when speed dropped below 20mph. It was not affected by weather or lighting conditions. The acoustic sensor was found to be a very economical option that was easily set up and performed well except in heavy rain and congested traffic. When speeds were over 40mph, count accuracy was within 10% and with slow speeds it was up to 32%. Speed accuracy was mostly within 5% except for lane 1 which overestimated speed by as much as 20 to 25mph during slow speeds.

2.2.5 Summary of traffic sensor evaluation

In the field tests described above, the complexity in selecting a preferred traffic sensor was described. Often the conclusion was that there are too many influences to be able to determine a single outcome, instead all traffic sensors providing reasonably accurate results should be evaluated on their own merits for the particular application. The acoustic sensors tested did not always provide the most accurate results, occasionally performing poorly in comparison with other technologies. Heavy rain sometimes adversely influenced results and low temperatures caused consistent under-counting in some cases. Nevertheless, the acoustic sensor was described in the evaluation studies as being an economical option with an acceptable accuracy. In summary, when rigorously tested against many other traffic sensing technologies, an existing acoustical traffic sensing product claimed a respectable performance under most conditions.

2.3 Discussion of traffic sensors

A range of different traffic sensors have been introduced and compared in this chapter. Some technologies such as video are more versatile, obtaining a larger range of parameters in a range of situations for many different applications. Other more traditional intrusive sensors like induction loops boast high accuracy at a lower cost. Each technology must tackle specific problems and has at least one disadvantage. In the case of infrared, atmospheric temperature and weather is a hindrance. Even with extensive and lengthy comparative tests as described in Section 2.2, the search for an optimal traffic sensor is unresolved. No single existing traffic sensor can provide a cost-effective solution in all applications with sufficient accuracy, reliability and flexibility.

It is possible to present a defence of audio-based traffic sensors against other technologies. Existing audio products such as the SmartSonic and SAS-1 described in Section 2.1.9 demonstrate that the market for and industrialisation of such a product exists in its own right. In the Texas Transportation Institute evaluation described in Section 2.2.4, the acoustic sensor was found to be a very economical option that was easily set up and performed well in heavy rain and congested traffic. When compared against video, audio has no difficulty with monitoring traffic in insufficient or changeable lighting conditions and visual occlusion is irrelevant. The large data storage and the computationally-hungry algorithms using video is unnecessary. Expensive cameras, sensitive mounting and calibration are not required. A microphone array also has the potential to monitor multiple lanes, as has already been demonstrated by the SmartSonic and SAS-1 products, matching the equivalent benefit of video. However, compared to a functioning induction loop or magnetic sensor, audio traffic monitoring system accuracy is most likely to be lower even in the most conducive environment for sound. Even so, installing and maintaining an induction loop is disruptive, time-consuming and expensive, especially in harsh weather conditions when roads are regularly damaged. Audio sensors are more versatile, economical and more beneficial if high accuracy is not critical.

Even if proven more advantageous, audio traffic sensors are never going to supersede all existing technologies. There are a variety of environments and purposes for which vehicular traffic requires monitoring, not all of which are appropriate for audio sensors. An inescapable reality is that audio sensors cannot measure the desired traffic sounds if an unrelated sound is overwhelming. Nevertheless, there is much to be gained from researching and developing audio traffic sensor techniques. By doing so, the boundaries of what is currently possible are marked, tested and sometimes moved. A fusion of data from different traffic sensing technologies may present more balanced, reliable, accurate and relevant information. In some situations such as a critically important junction or motorway with heavy traffic, it is worth investing in multiple complimentary traffic sensors. However, for a sparsely populated rural road with light traffic, such a system would be economically unviable and sensor data fusion is excessive. Moreover, until the quality and technology behind audio traffic sensor data is critically developed, there is little to be gained by fusing it with other sources. In conclusion, the research and development of audio traffic sensors is considered to be justified once the limitations are taken into account.

2.4 Conclusions

The survey of related vehicle sensing technology has shown that induction loops continue to be widely used due to their relatively low cost, familiarity and maturity. Non-intrusive sensors present a viable alternative, especially for multi-lane applications or situations where induction loops cannot be installed. However, no single existing traffic sensor can provide a cost-effective solution in all applications with sufficient accuracy, reliability and flexibility. Depending on the scenario, each sensor has its individual merits and application. Existing passive acoustic sensors present an economical and versatile option that have been proven to function well as traffic monitoring devices. While accuracy and fidelity may not be as high when compared with other sensors, it may be that the information obtained is sufficiently detailed for many applications.

Chapter 3

Road Acoustics

Understanding traffic noise characteristics and changes over time is central to developing an audio traffic sensor system. Therefore, it is necessary to gain some knowledge of vehicle noise as well as outdoor environmental issues before investigating an audio-based traffic monitoring approach. This chapter is concerned with the generation and propagation of traffic noise outdoors. Noise sources in a car are described under a variety of conditions, and compared against other vehicles. Since outdoor sound propagation heavily influences the information captured by acoustical sensors, some relevant background information on outdoor acoustics is introduced and discussed. Section 3.5 summarizes the chapter, drawing conclusions that are used in later experiments.

3.1 Sound characteristics

Sensors measure the presence or variation of a signal or stimulus, where that signal has particular attributes. Audio traffic sensors are heavily influenced by, and must operate within the constraints set by sound properties. To enable a discussion and understanding of such properties, this section presents some pertinent background information about sound.

Sound is generated when a vibration or oscillation causes a portion of a medium such as air to be displaced, generating an elastic force in the adjacent molecules. This displacement propagates longitudinally through the air with a finite speed (c) that depends on air properties such as elasticity and density. Detected as a pressure change at a particular frequency, this wave has a small magnitude[1] in comparison with the ambient atmospheric pressure [193].

Frequency (f) is the number of complete oscillations that a sound source undergoes per second. The frequency range of human hearing is usually described as being between 20 and 20,000 Hertz (Hz). The upper limit decreases with age and both limits differ from person to person. Described as infrasonic, frequencies below 20 Hz are felt even if they are not heard and constitute part of the overall sensory experience. Wavelength (λ) is the distance sound travels during each cycle of a sound source that executes repetitive motion, or is simply the distance between successive compressions or rarefactions. Frequency and wavelength are related by

$$c = f\lambda, \tag{3.1}$$

where c is the velocity of sound propagation in metres per second. The speed of sound c depends on the propagating medium. From the ideal gas law [49, 6], the speed of sound in an *ideal* gas depends on the type of gas and temperature, and appears to be independent of changes in pressure. In general, the speed of sound in air can be approximated as

$$c = 331.45 + 0.6T \text{ m/s}, \tag{3.2}$$

where T is in ^{o}C. Temperature dependence is one of the causes for the bending of sound waves, which can significantly affect propagation over long distances. Accurately estimating the speed of sound is important when modelling the location of a vehicle via inter-microphone time-delay in Chapter 6. For this reason it is beneficial to determine local temperature and also humidity and wind velocity when performing acoustical measurements.

Although pressure is measured in Pascals, sound level is customarily specified in *decibels*. It is a logarithmic scale suitable for human hearing (also logarithmic in behaviour) where the large dynamic range of human hearing is catered for. It can be shown that the instantaneous acoustic power associated with a sound wave is proportional to the square of the instantaneous pressure associated with an acoustic

[1] The sound pressure magnitude is generally in the range from 2×10^{-5}Pa to 20Pa (0-120dB) as compared with the standard atmospheric pressure of 101 325 Pa. The unit of pressure called the Pascal is equal to $1N/m^2$, named after the French mathematician and physicist, Blaise Pascal (1623-1662).

wavefront, or its mean square value p^2 [21]. With a reference pressure p_{ref} of 2×10^{-5} Pa, 0 dB corresponds closely to the threshold of hearing at 1kHz. Using the root-mean-square (rms) values of pressure, sound pressure level (SPL) can be written as

$$SPL = 20log_{10}\frac{p}{p_{ref}}dB, \tag{3.3}$$

where the reference pressure p_{ref} for airborne acoustic measurements is $2 \times 10^{-5} N/m^2$. Sound pressure is not to be confused with sound intensity L_I. The sound intensity is defined as the sound power level per unit area, whereas the sound pressure is the force per unit area. $L_I = \frac{p^2}{\rho c}$, where ρ is the density of air and c is the speed of sound.

Sounds, other than synthetically generated tones, typically contain multiple frequencies consisting of complicated repetitive waveforms which can be constructed from a Fourier series of harmonically related sinusoids, each with the appropriate amplitude and phase. Noise has a random nature, is not repetitive and contains all possible frequencies in a given range. By measuring the signal level in a series of frequency bands over a sufficient amount of time, a *frequency spectrum* of the sound can be obtained. Section 3.3.1 provides information on vehicle frequency spectra. The dependence of sound propagation on the atmosphere and particularly temperature implies that highly accurate assumptions about the source signal can only be made if pertinent atmosphere-related information is included, as the received signal may differ in frequency characteristics from the assumed signal. The sound pressure level and exact frequency spectrum is not directly relevant to the implemented traffic monitoring method. This reduces the need for accurate medium-related information.

3.2 Outdoor sound propagation

Propagation of noise in an open area is governed by a number of phenomena that adversely distort the source signal over distance. These phenomena set constraints on sensor placement. Relevant governing principles include geometrical spreading, atmospheric absorption, ground effects and refraction produced by vertical gradients of wind and temperature. Sections 3.2.1 to 3.2.4 describe the topics relevant to this work. It is assumed that the receiver is far away from the sound source, so the model of an omnidirectional point source can be used. At distances as short as 15m, ground effects, gradients of wind and temperature and their fluctuations all need to be taken into account [118].

3.2.1 Geometrical spreading

Geometric spreading is a result of the expansion of sound wavefronts radiating from a sound source. It is independent of frequency and has a major effect in almost all sound propagation situations. The wavefront at the receiver is typically classified as being *planar* or *spherical*, depending on the geometric spreading of the sound during propagation. Planar and spherical wave conditions are generated by *far-field* and *near-field* scenarios respectively. The far-field is defined as the sound field being sufficiently distant from the source so that the particle velocity is primarily in the direction of the sound wave [193]. In this work, the sound received at microphones is assumed to be a far-field planar wave, due to the choice of system parameters as described in Section 6.2.1.

Consider an ideal point source radiating spherical waves in Figure 3.1. As sound radiates spherically from an idealized point source s, the sound intensity level at r is related to the sound power level of the source by $\frac{1}{r^2}$[2][164]. The intensity of a sound wave is proportional to the sound pressure squared. By doubling the sound pressure, the intensity is quadrupled. Conversely, the attenuation of the rms sound pressure level is related to r from a source point by $\frac{1}{r}$, known as the inverse-distance law:

$$p = k\frac{1}{r}e^{-\alpha r}, \qquad \alpha = \alpha_1 + \alpha_2 \tag{3.4}$$

where k is a constant and α is a frequency-dependent sound attenuation coefficient for atmospheric absorption[3] described shortly in Section 3.2.2. Sound propagation losses due to spreading are normally expressed in terms of dB per doubling of r. For example, the sound level is reduced by 6 dB for each doubling of distance from the source for spherical waves [193].

The inverse square law is not the only cause of sound attenuation. If it were, then it would be possible to detect the sound of an aircraft at a distance of 100 miles. Since air is not a perfect lossless (perfectly elastic) medium, some sound attenuation

[2] This may be understood as a given amount of sound power being distributed over the surface of an expanding sphere with area $4\pi r^2$. Thus the sound intensity at a point $L_I \propto \frac{1}{r^2}$.

[3] α_1 is due to viscosity, heat conductivity and energy dissipation due to rotational energy states of air molecules, known as classical absorption. It can be neglected except at very high frequencies. α_2 is dominant and is a result of a complex molecular relaxation absorption. It is frequency, temperature and humidity dependent.

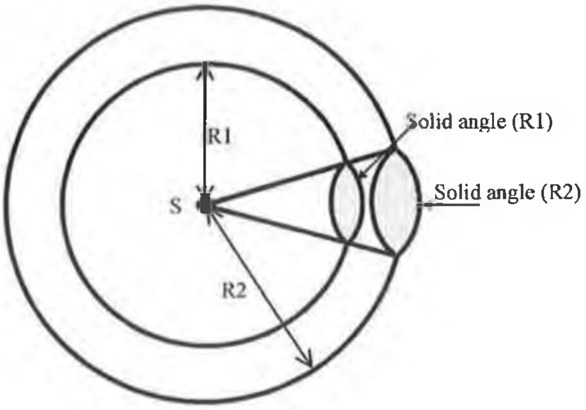


Figure 3.1: Sound attenuation over distance from a point source

must be attributed to absorption, as represented in Equation 3.4 by the frequency-dependent parameter α. Therefore, geometric spreading and attenuation impose a limit on the maximum source-receiver distance in measurements, for which the source sound can be detected by a microphone.

3.2.2 Atmospheric absorption

Atmospheric absorption depends upon frequency and relative humidity, and to a lesser extent upon temperature, since air molecules behave differently as these parameters change. The dissipation of sound energy due to atmospheric absorption is due to two major mechanisms: molecular relaxation (α_2) and viscosity effects (α_1), of which the most important by far is molecular relaxation. Viscosity effects are due to friction between air molecules which results in heat generation, known as *classical absorption*. Molecular relaxation absorption is where sound energy is momentarily absorbed in the air molecules and causes the molecules to vibrate and rotate. These molecules can then re-radiate sound at a later instant which can partially interfere with the incoming sound. Sound attenuation due to atmospheric absorption has been extensively studied and quantified in the international standard ISO 9613-1:1996[1]. From this source, Equation 3.5 is a basic expression describing a pure-tone sound propagating through the atmosphere over a distance r. The sound pressure amplitude p_t decreases exponentially as a result of the atmospheric absorption effects from its initial value p_i, in accordance with the decay formula for plane sound waves in free space.

$$p_t = p_i e^{-0.1151\alpha r}, \tag{3.5}$$

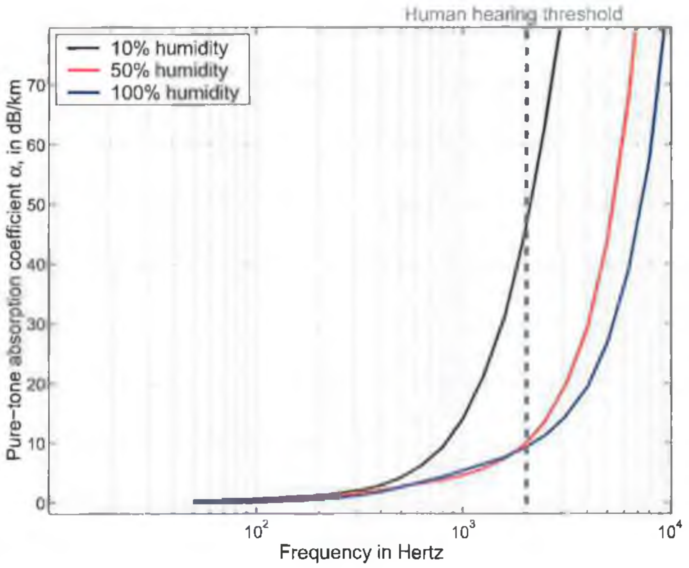


Figure 3.2: Air absorption coefficient α at 20°C, air pressure of 101,325kPa [1]

where α (also described in Equation 3.4) is a frequency-dependent sound attenuation coefficient for atmospheric absorption. Figure 3.2 illustrates values for α under different conditions. It can be seen that air absorption is substantial at higher frequencies, particularly the ultrasonic range. Also, as humidity decreases, attenuation increases. Although these values for α are based on a pure tone, while traffic noise is wideband, they provide a means to quantify the potential attenuation of sound due to air absorption and are sufficient for the approach taken in this work. Air absorption is relevant only over distances greater than a few hundred meters or at high frequencies [118]. It is important to note that the highest frequencies will be attenuated much more significantly than lower frequencies. Depending on the source-receiver distance, the frequency spectrum of the received signal may deviate significantly from the source frequency spectrum, particularly at higher frequencies. It should be reiterated that the received frequency spectrum is not always an accurate measure of the source spectrum.

3.2.3 Ground effects

Ground effects can cause attenuation at lower frequencies (200-800Hz) and have two components: *interference* and *impedance*. When both source and receiver are close to the ground, there can be interference between the direct and reflected waves, shown in Figure 3.3(a). The reflected portion of the wave leaves the surface at the angle of incidence, the amplitude and phase having been modified by the acoustical impedance of the surface. The direct and reflected waves merge at the receiver in a way that depends on their relative phase and amplitude, as this is a function of source height, distance and receiver height as well as the ground properties.

Ground effects have been described in the literature [165, 194, 42, 54, 58], with formulae that approximate the impedance effect of the ground under consideration [8, 9]. If the ground is absorbent, as is the case with grass and other foliage, there is an appreciable attenuation of the level of reflected sound. Thick grass may result in reflected sound levels being reduced by up to about 10 dB per 100 meters at 2000 Hz where high frequencies are generally attenuated more than low frequencies. A typical road surface has an effectively infinite acoustic impedance[4] for frequencies up to about 3000 Hz, according to Malherbe and Bruyère [119]. It can therefore be assumed that the noise source moves over a perfectly reflecting ground plane, which leads to interference between the direct and reflected sound waves at the receiver. Jonasson [86] describes problems encountered with microphone elevation and interference, whereby a height of 1.2m yields substantial sound prediction errors at 250Hz and above. With a lower microphone position, the problems move upwards in frequency but raise issues with ground attenuation.

A higher elevation may reduce ground effects and increase accuracy, but such a requirement places constraints on the mounting and potential suitability of such a system. Therefore it is preferable to develop a system that can tolerate some level of ground effects. A range of microphone elevations were used during experiments described in Chapter 8, from 1.5m to 3m. Promising results were achieved with the implemented time-delay method, based on these elevations.

3.2.4 Refraction from wind and temperature effects

Both wind and temperature variations in the atmosphere affect the energy distribution of sound by refracting the sound rays from their normal path [155, 107, 74]. Wind speed usually increases with height above the ground. As a result, the upper and lower part of the wavefront are affected differently, causing a bending or curva-

[4] The acoustic impedance of a material is defined as the product of density and acoustic velocity of that material. It is somewhat analogous to electrical impedance and is useful in assessing the absorption of sound in a material.

ture of the sound wavefront towards the ground in the wind direction, illustrated in Figure 3.3(b). The wind speed, in the absence of turbulence, typically varies logarithmically up to a height of 30 to 100 meters, then negligibly thereafter. Since wind speeds are much less than the speed of sound, a constant wind will have very little effect on the propagation of sound.

Temperature differences between the ground surface and air have a similar refractive effect [71]. Temperature usually decreases with increasing altitude, causing an upward curvature since sound velocity decreases with a temperature decrease. However a temperature inversion can occur, bending the sound towards the ground. The combined effects of temperature and wind gradients can result in measured sound level variations being as great as 20 dB. These effects are particularly important where sound is propagating over distances greater than a few hundred meters [80]. During experiments described in Chapter 8 the distance between the source traffic signals and microphones range from 0.5 to 8 meters. Refraction from temperature and wind is therefore minimal for sound propagation distances in experiments performed during this work. Furthermore, a precise measurement of the sound pressure level at the source is less important to the traffic monitoring system than an accurate measurement of the speed of sound propagation. Therefore, refraction from temperature and wind is not considered in the traffic monitoring system due to the measurement distance being less than 8 meters.

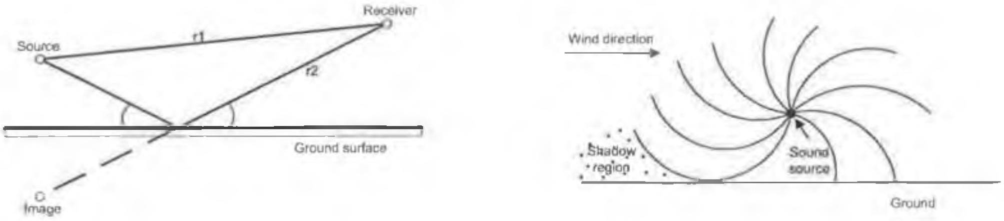


Figure 3.3: (a) Geometrical illustration of ground plane direct and reflected sound sources propagating to the receiver (b) Illustration of sound curvature towards the ground in the wind direction, due to increased wind speed with higher elevation

3.2.5 Cloud, fog and smoke

Precipitation, rain, snow, or fog have an insignificant effect on sound levels, although the presence of precipitation will affect the humidity and may also affect wind and temperature gradients. Wet road surfaces on the other hand, do affect the generated sound, as discussed in Section 3.3.2. Attenuation due to fog and smoke is mostly attributed to molecular absorption. However, if the particle size is very small, at low frequencies the particles can move with a velocity that approaches that of air molecules, causing a slight additional absorption [193]. The amount of additional absorption depends on particle size, species of smoke and frequency.

Sound travelling close to the ground will be attenuated by shrubs, bushes, leaves and trees as well as the soil itself. According to [193] a 100ft wide strip of foliage substantially reduces high frequencies, however it only reduces low frequencies by about 2dBA. Depending on the density and surface area of the foliage, attenuation of up to 30dBA may be achieved at 4kHz [13].

During experiments, it was not possible to find a suitable location to install a permanent recording system in close proximity to a road. This restricted the variety of conditions in which our traffic data was gathered. Therefore the effect of cloud, fog and smoke can only be estimated based on acoustics theory and research publications. In Section 2.2.4, a commercial audio beamforming-based traffic monitoring system was found to perform well in adverse weather, including rain. This indicates that the developed audio-based traffic monitoring system can perform well in a variety of weather conditions, though of course it is not conclusive proof.

3.2.6 Doppler effect

If either source or receiver is moving, then the frequency of the perceived sound may differ from that emitted. This is known as the Doppler shift. When the source and receiver are moving towards each other there is a rise in frequency, while if they are moving apart the frequency is reduced. Illustrated in Figure 3.4, it is shown in [193] that the perceived frequency f' is related to the emitted frequency f_s through the expression

$$f' = \frac{c}{c \pm v_s} f_s, \tag{3.6}$$

where v_s is the source velocity and has a negative sign when approaching the receiver.

The simplest sound wave is a continuous pure tone of fixed single frequency, however it rarely occurs outside the laboratory. The change in frequency has no noticeable effect on the sound pressure level of the source.

Couvreur and Bresler [45] used Doppler-based motion estimation for wide-band sources from single passive sensor measurements. Since only a single sensor was used, the approach involved the analysis of the acoustic signature to determine source speed and position. An ARMA [65] spectral estimator was utilized and the measured Doppler shift used to estimate source motion. Poor performance is reported due to background noise, inappropriate stationarity point source assumption and inadequate modelling of sound propagation effects. The reported poor performance of the Doppler-based motion estimation approach, combined with the lack of robustness of the method to background noise were two primary reasons for not utilizing the Doppler effect to detect moving vehicles in this thesis.

3.2.7 Summary of outdoor sound propagation effects relevant to traffic monitoring

This section has introduced the most relevant issues in outdoor sound propagation. Some relevant conclusions from this section that influence this work are summarized as follows:

1. Since air temperature has a significant impact on sound speed, any sound velocity-dependent measurements should take current temperature into account;

2. The effects of wind and temperature gradients may be ignored under normal

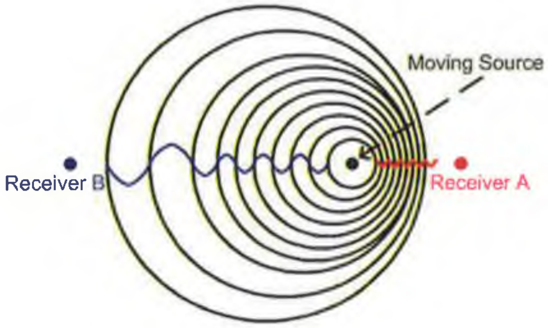


Figure 3.4: Doppler shift of the perceived sound frequency

conditions, provided the distance between source and receiver is within a hundred meters;

3. Ground effects become particularly relevant when the receiver is close to the road surface, suggesting a minimum and optimal elevation to maximize the capture of direct sound waves;

4. By making a far-field sound source assumption, planar wave propagation characteristics can be used;

5. The receiver should be close enough to the source to receive a signal that is not overly diminished in amplitude by the inverse square law and air absorption.

3.3 Road traffic noise

Road traffic noise is a wide-band sound signal generated by a variety of vehicles. These include cars, motorcycles and scooters, heavy vehicles such as trucks, lorries and busses as well as emergency vehicles such as ambulances, fire engines and police cars. Other sound sources from vehicles include the car horn, burglar alarm, ice cream van melody, and more lately, “boom boxes”. This section describes the sources of noise generated by a road vehicle as well as relevant influencing factors. The temporal and spectral pattern of a passing vehicle is described in Section 3.3.3 and standardized measurement procedures for measuring vehicle noise are mentioned in Section 3.3.3. Since transportation noise picked up on a highway may include passing trains and airplanes, Section 3.4 describes noise generated by other forms of transportation not encountered on a road, with a particular emphasis on how they differ from road vehicles.

3.3.1 Single vehicle noise

The generation of noise by a motor vehicle arises from a number of sources: the power unit (engine, exhaust, intake), cooling fan, transmission (gearbox and rear axle) rolling noise (aerodynamic and tyre/road interaction), brakes, body rattles and load [191]. These are commonly grouped into two categories; sources related to the power unit and transmission are referred to as *power train noise*, and all other sources are termed rolling or *tyre/road noise*. The relative importance of these sources depends on the operating conditions as well as the type of vehicle. Aerodynamic noise sources are not as important for exterior vehicle noise within legal speed limits due to the effective aerodynamic design that is necessary to meet fuel consumption requirements. An illustrative example of typical noise contributions is shown in Figure 3.5, based on measurements from different sources built after 1996 [169]. These were obtained in conformance with ISO 362 [3], the standardized method for measuring the noise emissions of individual vehicles, described in greater detail in Section 3.3.3.

Some industrialized countries introduced regulations to limit the maximum permissible noise emissions of road vehicles during the 1970s. Since their introduction, legal noise emission limits in the EU, Japan and US have been substantially lowered by as much as 16dB, depending on the vehicle type. The change in vehicle noise limits for passenger cars over time are shown in Figure 3.6. The current maximum level for a passenger car is around 76dB when measured in conformance with ISO 362, depending on the country. A substantial reduction of the power train noise emitted by cars has been achieved. This reduction is due to the encouragement of the aforementioned legislation as well as market research and technical progress. As engines and vehicle chassis become quieter, the power train noise becomes more or less equivalent to tyre/road noise for many vehicles. This results in tyre noise increasing in relevance, since it is the main contributor to vehicle noise during most driving conditions at constant speeds. For this reason, there is now a greater focus on reducing tyre/road noise in order to minimize noise pollution in developed countries.

Power train noise depends mainly on the engine rotational speed and the engine load, and is relatively independent of vehicle speed. Tyre/road noise starts to dominate over power unit noise at a certain *crossover speed.* This crossover speed depends on the type of vehicle, load and year of manufacture. Examples are shown in Table 3.1. A graphical illustration of the crossover speed is shown in Figure 3.7, where

Table 3.1: Crossover speed between power train and tyre/road noise [169]

Vehicle type	Cruising	Accelerating
Cars 1985-95	30-35 km/h	45-50km/h
Cars 1996 -	15-25 km/h	30-45 km/h
Heavy vehicles 1985-95	40-50 km/h	50-55 km/h
Heavy vehicles 1996 -	30-35 km/h	45-50 km/h

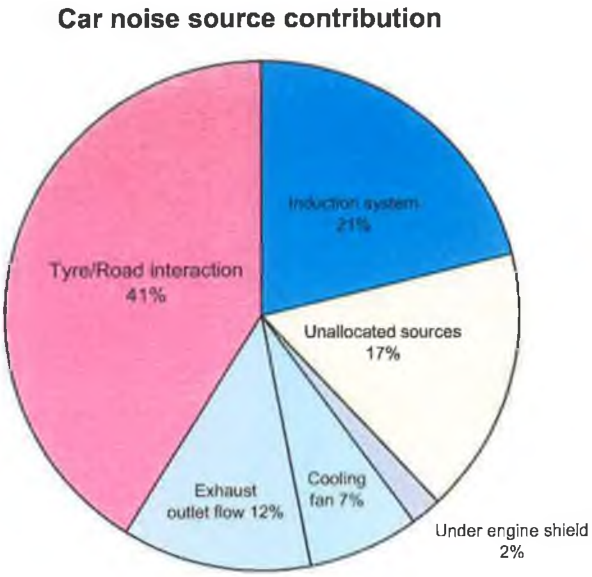


Figure 3.5: Distribution of car noise sources [169]

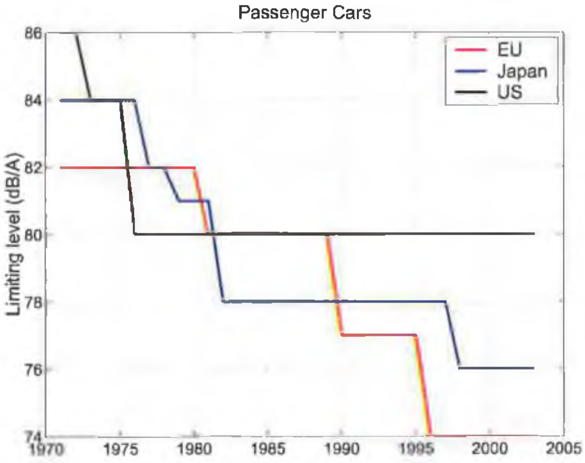


Figure 3.6: Development of legal vehicle noise emission limits over 25 years [169]

the noise level of a large amount of passenger cars travelling at a variety of speeds is shown. The noise level is separated into power train and tyre/road noise. The balance between these two noise sources at different velocities can be observed. In the graph, the crossover frequency can be estimated at 20-25 km/h, above which the tyre/road noise dominates.

3.3.2 Tyre/road noise

Tyre/road noise level and characteristics depend on a large range of parameters, not least vehicle velocity. There is much ongoing research on methods to reduce tyre/road noise, by changing tyres and road surfaces alike. Since transportation noise is arguably the predominant outdoor environmental noise pollutant, significant efforts in reducing such noise are highly relevant, making tyre/road noise a topic of study since the 1970s. It is foreseen that tyre/road noise will be reduced at some point in the future, either by tyre or road surface changes.

There is an extremely complicated mix of mechanisms and related phenomena that have some influence on tyre/road noise. It is not intended to investigate tyre/road noise generation here beyond a basic understanding of the resultant noise characteristics, since to do so would be outside the scope of this work. Furthermore, the approach taken in this work deliberately does not require a precise measurement of the absolute sound level or the sound characteristics. As a consequence the generated noise and future trends are described, as opposed to an in-depth study of sound source generation.

Tyre/road noise generation mechanisms can be divided into two main groups: structure-borne *mechanical vibrations* and air-borne *aerodynamic phenomena*. The noise is influenced by longitudinal forces (acceleration or braking) as well as by tangential forces (cornering) acting on the tyres. Also relevant are amplification/absorption effects and sound directivity. Examples of mechanical vibrations include the impact of tyre tread blocks on road surfaces, the effect of road surface texture on tyre tread, relative motion between the rubber and the road and temporary adhesion of the rubber to the road. Aerodynamic displacement includes air pumping in and out of cavities in or between the tyre tread and road surface, resonances in the tyre tread grooves and Helmholtz resonances[5] between connected air cavities. The ex-

[5] A Helmholtz resonator is an air cavity with an opening. A body of air in and near the open

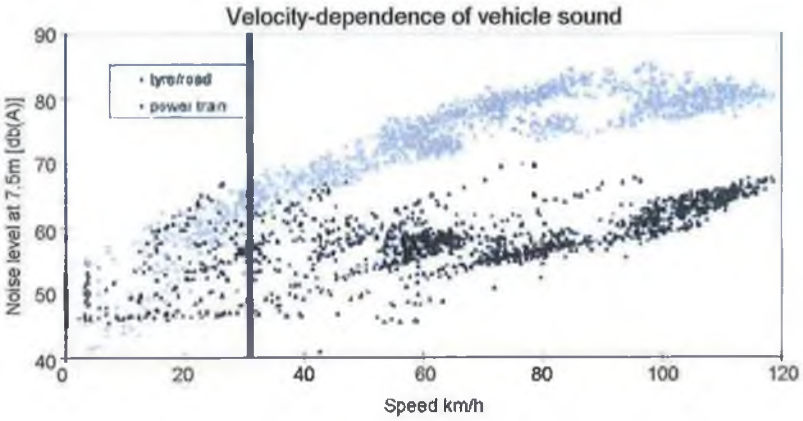


Figure 3.7: Car noise sources at different velocities with a crossover speed 20-30km/h [81]

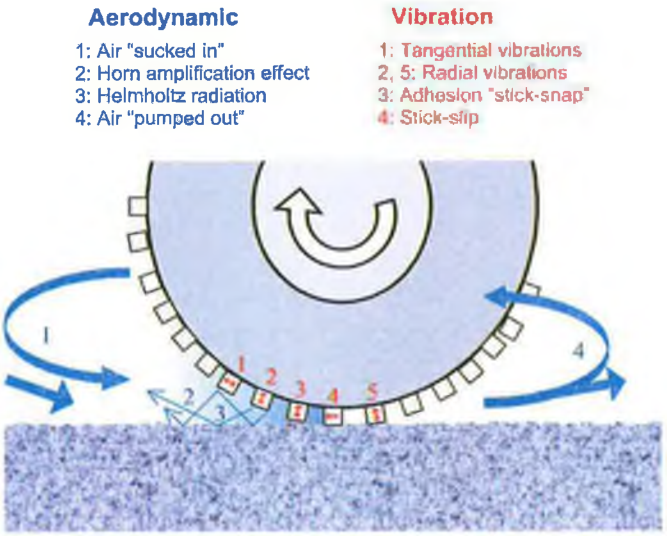


Figure 3.8: Noise sources due to tyre/road interaction [169]

ponential horn shape between the tyre and road surface has an acoustical matching effect, while porous surfaces on roads act like sound absorbing material. A subset of tyre/road noise generating phenomena is illustrated in Figure 3.8. Table 3.2 provides an estimate of the level at which major factors may influence tyre/road noise [169]. It can be seen from Table 3.2 that vehicle speed, and to a lesser extent road and tyre type, highly influence the overall sound. Studies comparing tyre noise give conflicting results on the range of sound levels between the noisiest and quietest tyres, many stating the range to be 3dB while others claim it is up to 9dB [169]. Truck tyres, carrying an estimated 10 times larger load than car tyres are on average 3-4dB noisier.

Table 3.2: Factors influencing tyre/road noise [169]

Speed	25 dB (30-130 km/h)
Road surface (incl. extremes)	17 dB
Road surface (conventional)	9 dB
Truck tyre type (conv., one size)	10 dB (same size)
Car tyre type (conventional)	8 dB (same width)
Studs in tyre (rel. to no studs)	8 dB (for new studs)
Load and inflation	5 dB (±25%)
Road condition (wet/dry)	5 dB(heavy rain)
Temperature	4 dB (0-40^{o}C)
Torque on the wheel (normal)	3 dB (0-3m/s^2 accel.)

The sources contributing significantly to the overall sound level are all located very low, in general within 50 or 100mm from the road surface. In principle, the entire tyre radiates sound, however the major sources are located at and very near to the leading and trailing edge of the tyre/road contact patch as well as at the tyre sidewall. In general the level of emission from the front of the tyre is slightly higher than from the rear. The body of the vehicle affects sound radiation substantially, especially in the vertical direction.

Sound *directivity* is another complicating feature of tyre/road noise that depends on the combination of tyre and road surface and source locations [169]. Horizontal directivity is substantial, where sound radiation is normally highest to the front, second highest to the rear and lowest in a direction perpendicular to the tyre rolling direction. Directivity is most pronounced on smooth-textured surfaces. Vertical

hole vibrates at a single resonant frequency because of the "springiness" of the air inside.

directivity is also substantial. This depends partly on the vehicle body screening effects, partly on the focussing due to the horn effect, and in general it means that sound radiation is lowest in an upward direction and highest at a rather low angle to the road surface.

Major studies have been undertaken in the past few decades exploring tyre sound generation mechanisms. Sandberg show that despite radical developments regarding safety and economy over the previous 60 years, tyre/road noise emission has been approximately constant, irrespective of tyre year model [168, 167, 169]. In summary, power unit noise has decreased, but tyre/road noise has remained the same and according to [167] has even increased in some cases.

Road surfaces

The road surface has an influence on the noise level, where the range between an extremely noisy surface and a quiet surface is approximately 17 dB [169]. Porous surfaces are generally less noisy than dense ones. With the same road surface, increasing chipping size generally means increased noise. Paving stone surfaces can be very noisy. The ISO 10844 standard [2] specifies the test track characteristics (as opposed to specific material) for measuring noise emitted by road vehicles. One of the requirements to conform with ISO 10844 is that the road surface must be no greater than the defined sound absorption coefficient α of 0.10. The ISO 10844 surface is one of the quietest surfaces, except for the porous surfaces. Road surface characteristics that affect tyre/road noise emission include the surface texture, porosity and layer thickness. The noise increase for a wet road surface is substantial at frequencies above 1kHz, but the effect on the overall levels is not high. Sandberg attempted to estimate the effects of wet surface on A-weighted[6] sound levels as shown in Table 3.3.

3.3.3 Measured road traffic sound characteristics

The temporal and frequency characteristics of traffic noise measured by a microphone adjacent to the road are now described. Traffic noise consists of a combination of

[6] A-weighting is a frequency-dependent weighting of sound signals, which has the greatest sensitivity in the 1 kHz to 5 kHz range. This corresponds to the range of the greatest sensitivity of the human ear and is the most common frequency weighting used for sound-level meters.

Table 3.3: Influence of a wet road surface on sound level [169]

Degree of moisture	0-60 km/h	61-80 km/h	81-130 km/h
Dry	ref	ref	ref
Humid	+ 2 dB	+ 1 dB	0 dB
Wet, moderate rain	+ 4 dB	+ 3 dB	+ 2 dB
Wet, intensive rain	+ 6 dB	+ 4 dB	+ 3 dB

multiple heterogeneous vehicles whose acoustical properties merge into one overall traffic sound. The level of highway traffic noise depends on the amount, general speed, and type of vehicles. The loudness of traffic noise is typically increased by higher quantities of traffic, higher speeds, and greater numbers of trucks. Since most vehicles produce very similar sounds, they can often be virtually indistinguishable and only identified as distinct vehicles by their temporal disparity. Heavy vehicles such as buses and trucks are generally louder than cars. Noise from heavy vehicles originates from the same vehicular components as cars. However, truck engines are used in the wide-open throttle mode for a greater portion of the time and in larger trucks the engines are more powerful, resulting in a greater sound intensity. The amplitude of sound from motorcycles is typically greater than for cars. The frequency spectrum of motorcycle sound contains stronger high-frequency components than a car.

Temporal changes

From the perspective of a roadside observer or sensor, the generated traffic noise is perceived as a series of passing sound sources. The noise level varies over time, with a peak when a vehicle is in close proximity to the sensor. In Figure 3.9 a time-frequency spectrogram of a single passing police car is shown, where the siren including its doppler shift is clearly visible as a sinewave at a low frequency. This temporal variation in sound level contains useful information that may be exploited, as is described in Section 7.1. When a vehicle passes by a roadside microphone, the sound signal increases as the vehicle approaches, reaches a maximum approximately when the vehicle is at its closest point to the microphone, and decreases as the vehicle passes away. There is an asymmetry of the signature that becomes more pronounced as the speed of the source increases, as described by Favre [60]. The maximum amplitude level is displaced to the right, with respect to the time at which the source is opposite the reception point. According to Favre, this asymmetry is due to the time of integration of the sound level plus the Doppler effect (the speed of propagation of the sound waves thus being taken into account). As a result of these two effects, the source appears to be downstream from its true position.

Standardized vehicular noise measurements

A standard procedure for measuring noise emitted by a passing road vehicle under urban traffic condition is specified in the ISO R362 standard [3]. The ISO R362 specifications are intended to reproduce the noise levels which are produced during the use of intermediate gears with full utilisation of the engine power available as may occur in urban traffic. ISO standard 10844 [2] specifies the test tracks to be used. The purpose of these specifications is to be able to determine the maximum noise a vehicle is capable of creating. During this type of driving, the engine develops maximum or close to maximum power and the resulting noise is dominated by power unit noise (engine, exhaust, transmission, air intake, fan etc.). Figure 3.11 shows appropriate test site dimensions, where the microphones are 7.5m from the centre of the road at an elevation of 1.2m. The specifications are intended to reproduce the noise levels that are produced during the use of intermediate gears with full utilization of the engine power available, as may occur in urban traffic. The vehicle approaches the test track at a constant speed. 10m before the microphone, the vehicle is accelerated with a wide-open throttle until it has passed 10m beyond the microphone, when the throttle is closed. The initial constant speed is generally 50km/h for all vehicles, with cars using 2nd and/or 3rd gear, and heavy vehicles using a wide selection of gears. The maximum noise level at the two microphones during the acceleration process is recorded and averaged over a series of repetition runs.

The European Committee for Standardization (CEN) has standardized a frequency spectrum for use in traffic noise calculations, based on typical frequency spectra of roadside traffic noise [5]. Shown in Figure 3.10, the EN1793-3 spectrum is intended to represent mixed light and heavy vehicle traffic in urban conditions, at a speed of around 50km/h. It is the same as in EN ISO 717-1, and is a quite useful verification of the broadband nature of vehicular sound.

The TÜV Süd test centre in Munich was visited by the author in order to obtain recordings of vehicles according to the ISO R362 standard. The test site is illustrated in Figure 3.12. The weather conditions at the time of recording included a tempera-

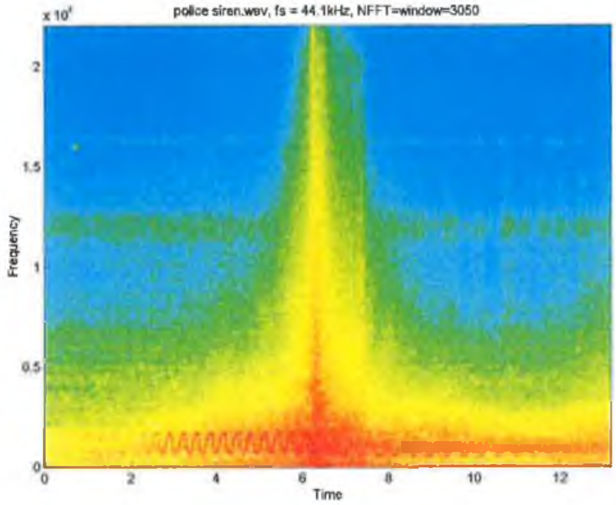


Figure 3.9: Spectrogram of police car. The Doppler-shifted siren is visible as the red oscillatory trace at the bottom of the spectrogram

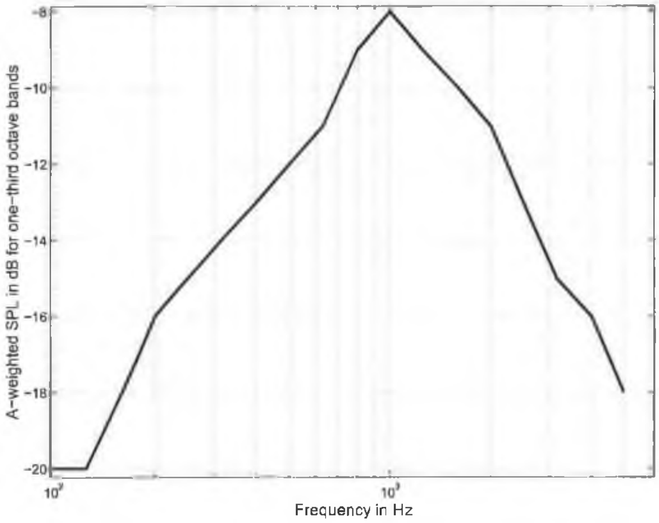


Figure 3.10: EN-1793 car noise frequency spectrum [5]

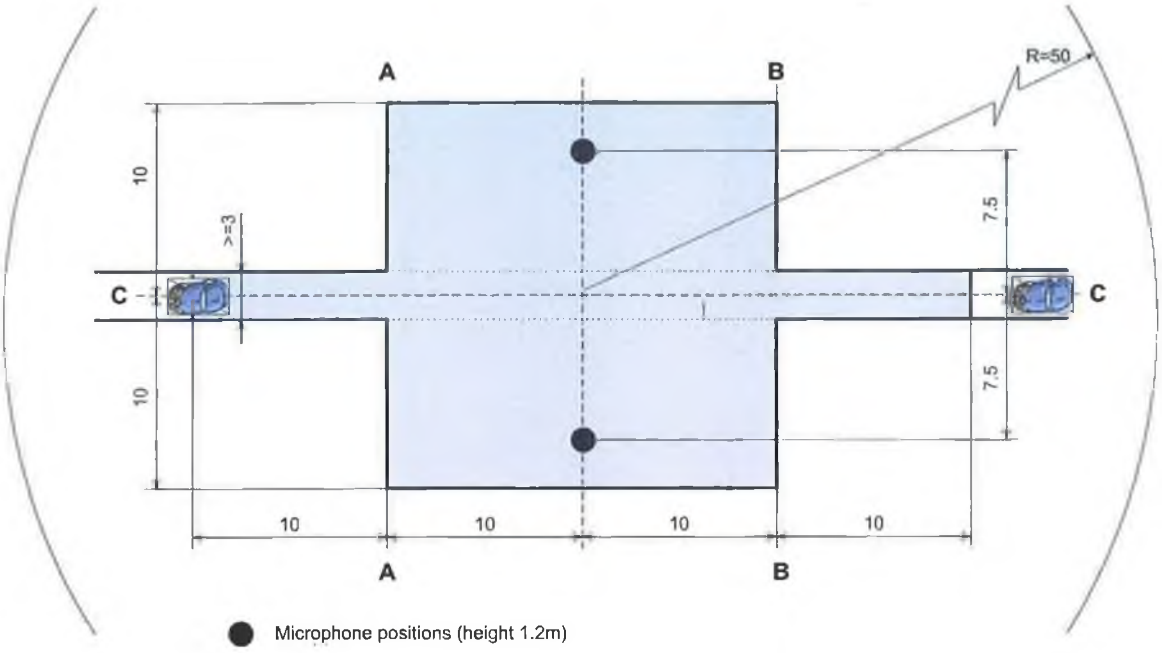


Figure 3.11: Geometry of an ISO 362 test site

Figure 3.12: Photograph of the ISO 362 test site in TÜV SÜD Test Centre, Munich

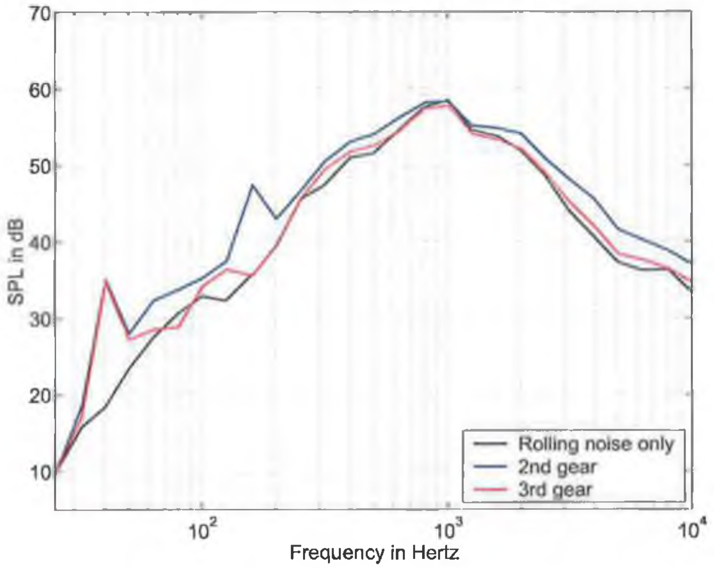


Figure 3.13: Frequency spectrum of a car in 2nd and 3rd gear and with no engine

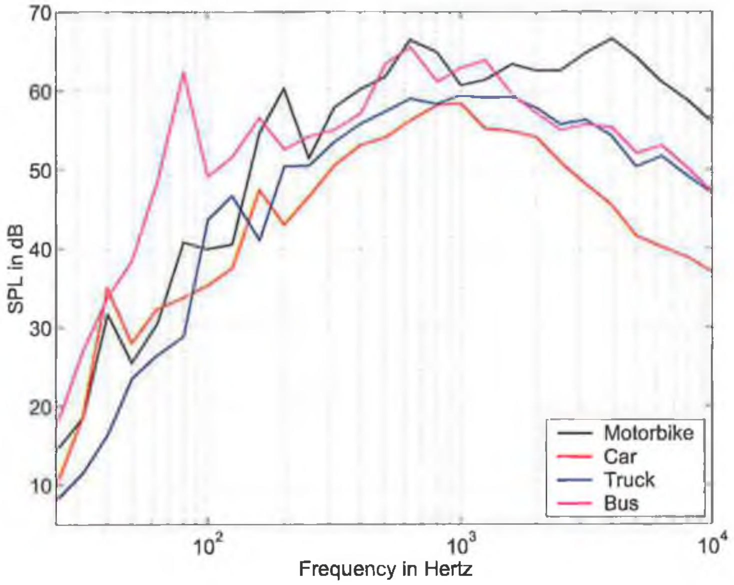


Figure 3.14: Frequency spectrum of different vehicle types measured according to ISO 362

ture of 19.6C, wind speed of 1.3m/s, air pressure of 7.4 hPa and a relative humidity of 53%. Frequency spectra were obtained for 61 cars. Each vehicle was recorded when travelling in 2nd gear, 3rd gear, and when rolling past the microphone with the engine switched off. Figure 3.13 presents the average frequency spectrum over all 61 vehicles for each type of recording. The only possible noise source for the spectrum when the engine is switched off is the tyre/road interaction. One interesting observation is that the difference in frequency spectrum magnitude between a vehicle with the engine switched on and off is very small. This confirms the statement that the tyre/road interaction is the dominant noise contribution for vehicles travelling above 30 km/h.

It can also be observed from Figures 3.9, 3.10 and 3.13 that the sound generated by a car typically consists of frequency components throughout most of the audible frequency range. A large portion of the energy is centered around 1kHz with a gentle roll-off to form a generally wide and flat frequency spectrum. There are no significant components. Unlike the engine, the tyre/road noise does not generate any harmonic tones. Since most of the sound generated by a moving vehicle is overwhelmingly the tyre/road noise, it can be assumed that the frequency spectra in Figure 3.13 are representative of most vehicles. Therefore the general frequency spectrum of a vehicle is typically a wide-band, flat noise-like spectrum without harmonic components.

The average frequency spectrum of each class of vehicle recorded at the TÜV Süd test centre is illustrated in Figure 3.14, to include a motorcycle, car, bus and truck. The wide-band shape of all frequency spectra are generally similar. Car noise has the lowest overall magnitude. Average motorcycle noise is quite high and particularly strong at the upper frequencies, especially when compared to other vehicle spectra. In addition to noise from intake, exhaust, and gearing systems, motorcycles radiate considerable noise directly through the engine walls. Exhaust noise is often sufficient to mask most other sound sources. Truck and bus frequency spectra are very similar above 2kHz, with a high overall magnitude relative to car noise. In general, all 4 categories of vehicles demonstrate similar spectral characteristics. This indicates that road vehicle classification based on frequency spectrum alone is a difficult challenge that may not present reliable results.

3.3.4 Summary of vehicular noise relevant to traffic monitoring

This section has described vehicular noise generation and sound characteristics to provide some measure of the possibilities and limitations regarding vehicle sound analysis. Relevant observations are summarized as follows:

1. The noise level and frequency spectrum of a vehicle is governed by a wide variety of parameters, from engine speed and road surface to environmental weather, background noise and receiver location;
2. Individual road vehicle noise is slowly getting quieter and is likely to continue to do so in the future as manufacturers minimise tyre/road noise. It is impossible to exactly define temporal-spectral vehicular noise characteristics, since these may change over time;
3. Moving traffic noise generates a broadband signal with a lack of perceptible dominant frequencies;
4. Vehicle and engine velocity have an impact on the characteristic traffic noise;
5. The frequency spectrum of vehicles does not differ significantly within vehicle class or from one class of vehicle to another, therefore it is difficult to classify a vehicle based on frequency spectrum alone.

3.4 Non-vehicular transportation noise

Transportation noise encompasses more than road vehicles, including airplanes, helicopters and trains among others. While not typically found on a highway, these sound sources may interfere with road vehicle monitoring as they pass. As such, it is useful to investigate characteristics of such sound sources if only to deliberately ignore them or classify them as background noises.

Aircraft noise has a unique frequency spectrum that is significantly different to road vehicular noise and can be categorized as turbojet aircraft, propeller fleet and helicopter noise. Takeoff, approach and landing of aircraft may lead to a noise of more than 100dB(A) at the ground, which may potentially mask road traffic noise.

Since aviation noise became a major public issue in the 1960s and 1970s, legislative controls have been brought in and quieter aircraft have been developed. Modern high-bypass turbofan engines, for example, are significantly quieter than the turbojets and low-bypass turbofans of the 1960s. Helicopters generate a very specific sound that is easily recognized. The acoustic signature is typically perceptible over a long time and contains strong frequency patterns, making it significantly different to vehicular sources. Train noise rarely reaches the amplitude of aircraft noise, but may nonetheless interfere with vehicular traffic sounds. Despite the wide variety in train types and noise sources, some common characteristics exist. Examples are the long pass-by duration due to train length and repetitive rhythmic sound often caused by wheel-rail interaction.

3.4.1 Turbojet aircraft

The noise produced by modern turbojet aircraft contains acoustical energy over a wide frequency range. The audible noise varies from a very low-frequency rumble to a very high frequency whine, depending on the aircraft type and the operation being performed (takeoff, landing, or ground run-up). Most of the sound energy from aircraft operations is found at lower frequencies. All aircraft engines are heat engines that convert rapidly expanding gas mostly into thrust, but a small portion is converted to sound waves.

Aircraft noise is generally divided into two sources: that due to the engines, and that associated with the airframe itself. As higher bypass ratio engines have become more common and aircraft have become larger, interest in airframe-related noise has grown, but engine noise still accounts for most of the aircraft external noise. A turbojet engine produces two kinds of noise: turbulence generated by the interaction of the high velocity jet with the stagnant atmosphere and a high intensity whine caused by the high speed rotation of the engine's multi-bladed fan-compressor. Aerodynamic noise arises from the external airflow around the aircraft fuselage and control surfaces. This type of noise increases with aircraft speed. It also increases at low altitudes due to the density of the air. Jet noise is a broadband noise source caused by the turbulent mixing of the high speed exhaust with the ambient air, where most of the energy is directed aft of the engine at a 45 degree angle from the engine axis. Turbo machinery noise often includes discrete tones associated with blade passage frequencies and their harmonics, as can be observed in Figures 3.15 and

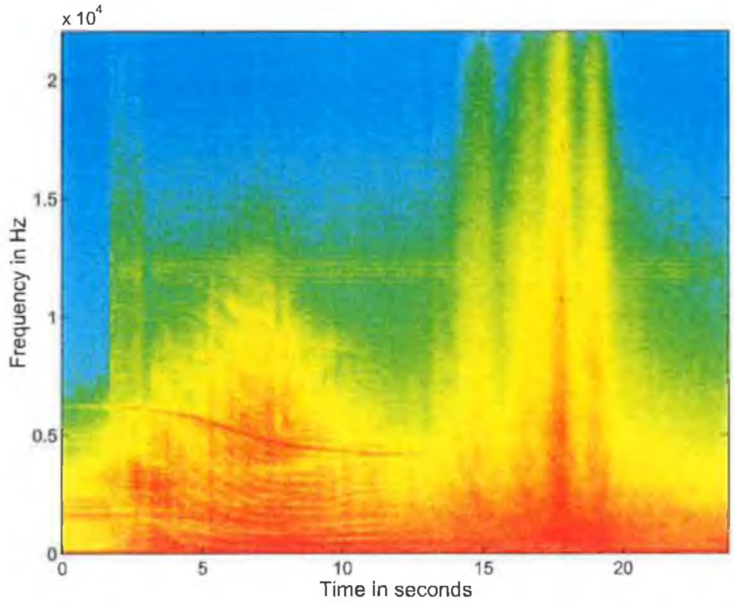


Figure 3.15: Spectrogram of a jet aircraft landing at Dublin airport followed by 4 cars in close succession between 15 and 20s. Recorded at an adjacent road with a sampling frequency of 44.1kHz and 5050 FFT samples

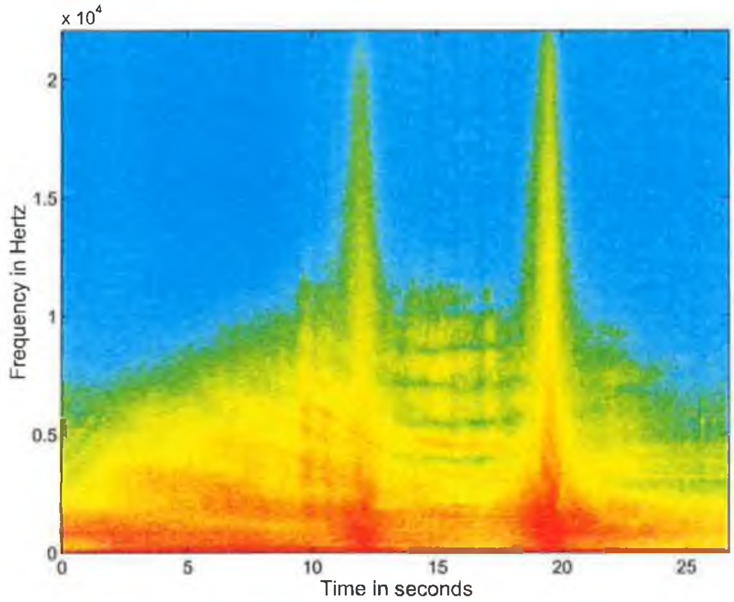


Figure 3.16: Spectrogram of another jet aircraft taking off at Dublin airport during which two road vehicles pass at 12s and 20s, visible as a sharp spike. Recorded at an adjacent road with a sampling frequency of 44.1kHz and 5050 FFT samples

3.16. These are from sounds recorded by a microphone pair at the side of a road running parallel to Dublin Airport runway. The spectrogram of a turbojet aircraft landing and taking off is shown, together with a number of cars passing. Over the past 30 years significant research has been conducted to reduce aircraft propulsive noise such that airframe noise has become a significant noise source for large aircraft during landing operations.

3.4.2 Propeller aircraft noise

Much of the noise of a propeller-driven aircraft is aerodynamic noise due to the flow of air around the propeller blades. Engine noise contributes to the general noise level in an aircraft. Propeller noise consists of (1) discrete frequency or rotational noise arising from periodic disturbances of the air by the propeller and (2) broadband or vortex noise arising from random disturbances at the propeller [163]. The discrete frequency noise results from pressure waves being generated by the rotating propeller blades, the frequency of oscillation corresponding to the blade-passing frequency and harmonics. The actual magnitude and waveform of the oscillating pressure depends on propeller design, rpm, thickness and thrust or torque forces on a blade element. Virtually all periodic propeller noise is low frequency. The broad band or vortex noise is produced by air turbulence in the wake of the propellers and by complex fluctuating forces that are exerted by the propellers on the air stream [193].

3.4.3 Helicopter noise

At a moderate distance from a helicopter, the primary noise sources have been identified as blade slap, piston or turbine engine exhaust noise, tail rotor rotational noise, main rotor 'vortex' noise, main rotor rotational noise, gear box noise, turbine engine noise and miscellaneous aerodynamically and mechanically produced sounds. Figure 3.17 illustrates a helicopter noise spectrum.

By the 1960s, the noise of helicopters had become an important issue. Initially, both the engine and the rotor were the major generators of noise. With the introduction of the turbo-shaft engine, the engine noise became less significant and the rotor became the dominant external source of noise. The main rotor and the tail rotor emit unique and recognizable sounds due to their highly individualized operating condition. The acoustic frequencies associated with the rotating blades are directly related to the

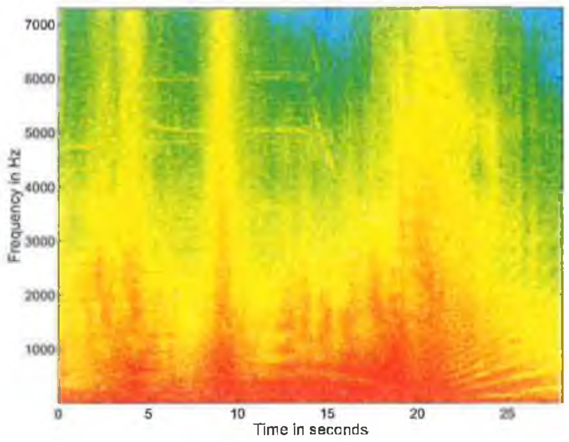


Figure 3.17: Spectrogram of a helicopter passing overhead

blade spacing. A helicopter main rotor generates primarily low frequency noise and, in certain operating regimes, high amplitude low-to-mid-frequency noise modulated at the blade passage frequency. The low frequency rotor noise is made up of basic loading noise and broadband turbulence noise, each a function of lift and rotational speed. These sources are present in any lifting rotor. Additional sources, such as Blade Vortex Interaction (BVI) noise and High Speed Impulsive (HSI) noise, become dominant in specific operating regimes, namely in descents and at high forward airspeeds, respectively. BVI noise can be the most significant contributor, because it occurs during a helicopters approach to the landing area.

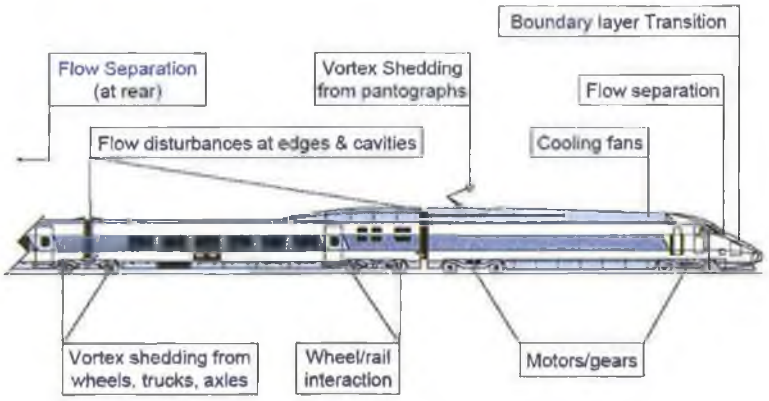

Boundary layer Transition
Flow Separation (at rear)
Vortex Shedding from pantographs
Flow separation
Flow disturbances at edges & cavities
Cooling fans
Vortex shedding from wheels, trucks, axles
Wheel/rail interaction
Motors/gears

Figure 3.18: Typical steel wheel high-speed train noise sources [4]

3.4.4 Train noise

A description of train noise is complicated by the wide variety of train types and operating conditions. Noise generated by a train on its surrounding environment is a function of a number of different factors including the interaction of the wheels and rails, the vehicle propulsion system, auxiliary equipment, noise radiated from vibrating structures, train speed, train length and aerodynamics [131]. As well as the airborne noise, ground-borne noise and vibration traveling through the track and support structure is experienced as a low-frequency rumbling noise or as a mechanical vibration. Railway noise depends heavily on the speed of the train, as is clearly illustrated in Figure 3.19.

Trains are traditionally associated with diesel or electric locomotives which push or pull either freight or passenger rail cars. In this case, the generated noise is generally characterized by a high noise level during the locomotive pass-by with lower noise levels or noises of different character as the carriages pass by. Electric self-propelled trains common in large urban areas have no locomotive. Maglev trains are magnetically levitated and powered high-speed systems representing the upper range of speed performance up to 300 mph. While the very high maximum speeds make maglev trains very attractive, the high cost of the lines has limited their current commercial application to one line in Shanghai [44].

The total noise generated by a high-speed train pass-by can be generalized into three major categories: propulsion noise, mechanical noise from wheel/rail interactions and/or vibrations, and aerodynamic noise resulting from airflow moving past the train [4]. For a conventional train with a maximum speed of up to about 125 mph, *propulsion* and mechanical noise such as those described in Figure 3.18 are the predominant sound sources. Fan noise tends to dominate the noise spectrum in the frequency bands near 1000 Hz. The spectrum for *wheel-rail interaction* rolling noise peaks in the 2 kHz to 4 kHz frequency range. It dominates the sound level at speeds up to about 160 mph and increases more rapidly with speed than does propulsion noise, typically following the relationship of 30 times the logarithm of train speed. Above 160 mph *aerodynamic noise* sources tend to dominate the radiated noise levels.

Regardless of train type, the duration and frequency spectrum of a passing train is significantly different to that of a vehicle. Although high-speed trains may travel

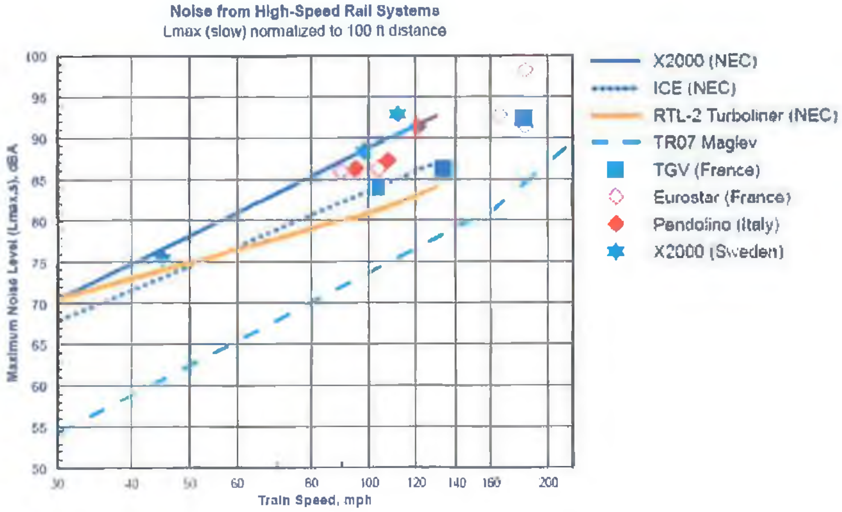


Figure 3.19: Measured values of $L_{max,s}$ vs speed from high-speed rail systems [4]

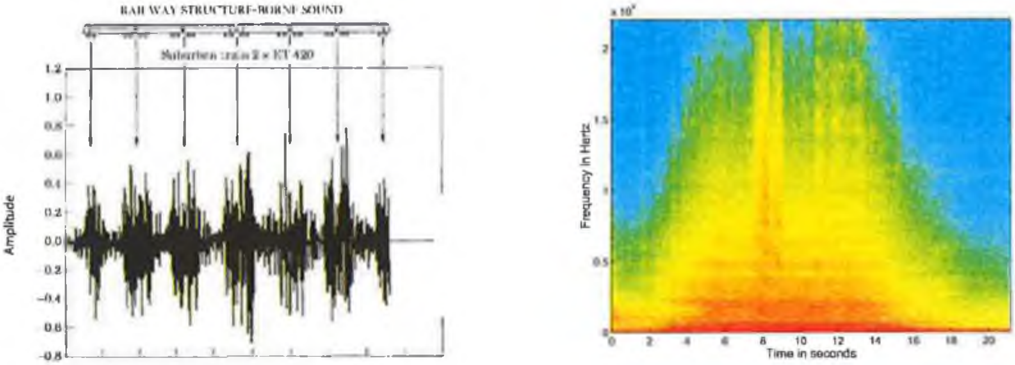


Figure 3.20: Passing Train (a)temporal sound [73] and (b) spectrogram (from Skerries train2.wav)

much faster than vehicles, their significantly larger length dictates a far longer pass-by duration. Figure 3.20(a) illustrates the sound characteristics of a passing train. Where the acoustical signature of a passing vehicle is typically less than 10 seconds, a train acoustical signature can be much longer. Depending on the circumstances, a rhythmical or repetitive sound can often be heard from a passing train. This is illustrated in Figure 3.20(a). These two temporal characteristics together with frequency spectrum characteristics are useful factors in indicating the passage of a train.

3.4.5 Summary of non-vehicular noise relevant to traffic monitoring

Some relevant conclusions from this section on non-vehicular noise that influence this work are summarized as follows:

- Aircraft noise may overwhelm traffic noise but is sufficiently different to be distinguishable as irrelevant to road traffic monitoring;

- Trains are closer to road vehicles in loudness, but like aircraft are significantly different as to be distinguishable.

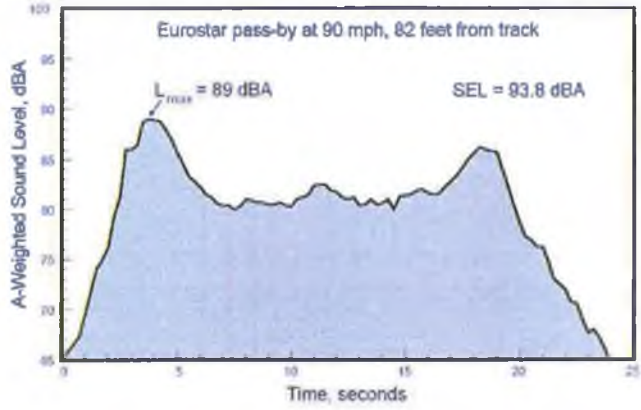


Figure 3.21: Eurostar pass-by noise [4]

3.5 Conclusions

This chapter has described the relevant aspects of outdoor sound propagation, transportation noise generation and measurements. Outdoor acoustical effects increase the difference between measured sound and characteristics of the sound originating at the source. Moreover, multiple sounds from heterogeneous vehicles are measured simultaneously, thereby increasing the difficulty in identifying a particular vehicle or its behaviour. The frequency properties of vehicular noise are generally that of a broad, wide-band, noise-like spectrum with a lack of perceptible dominant frequencies. Sometimes other sounds may interfere, such as non-vehicular transportation. However, the characteristics and velocity of other sounds are sufficiently different as to avoid mis-classification. Chapter 6 describes the derivation a model for a moving sound source along a particular trajectory that increases the system ability to distinguish between vehicles travelling along the road being monitored and other locations.

Chapter 4

Sound Source Localization

If the only problem was to count cars that were suitably spaced out on a good road, simple analysis techniques could easily be implemented. Unfortunately, road vehicles do not have a homogeneous type, velocity or spacing. A side-firing microphone array may need to distinguish between multiple distributed vehicles sources as well as determine relevant characteristics of each vehicle. The question at hand is therefore how many sources are present and what are their locations and characteristics? Source localization techniques seek to resolve this question by determining the spatial location of a source based on multiple observations of the emitted sound signal.

This chapter describes relevant sound source localization techniques that may be applied to determine the presence and location of road vehicles. Section 4.5 summarises the reasons for choosing a cross correlation method. Chapter 5 details the implementation of a time-delay of arrival (TDOA) cross-correlation approach.

4.1 Background information

In sound source localization, the desired information is the position of the sound emitting source - the acoustical characteristics are largely irrelevant. A minimum of two or more spatially distributed sensors are required to determine the location of a source. Arrays of two or more sensors are often used to increase accuracy. The purpose of a sensor network is to monitor an area; detecting, identifying, localizing and tracking one or more objects of interest. There are a choice of established techniques, some of which date back to around World War II. The choice of method depends on a number of factors, some of which are listed in Table 4.1.

Table 4.1: Factors affecting the choice of sound source localization method

Source propagation	number of sound sources type of source knowledge of propagation speed environmental reverberance
System Geometry	number of microphones relative microphone placement array geometry knowledge of source-receiver geometry
System specifications	required accuracy computational power sensor synchronization

Accurate source localization estimation is of fundamental importance in many applications, e.g. intelligent living environments, speech separation for hands-free communication devices, security systems, teleconferencing and acoustic surveillance systems. Transmitted information used for localization may be in the form of sound or electromagnetic waves. Radio frequency (RF) electromagnetic waves are used by wireless devices to determine their position based on either signal strength [192], time of arrival [180, 41], angle of arrival [189] or a hybrid of signal strength and time. RF signals can be applied to indoor and outdoor non line-of-sight scenarios over a larger distance than audio. Vision-based localization using optical sensors is limited to the visible surrounding environment and is particularly relevant to robot technology [186, 40]. Approaches using vision can use 3-D maps of the surrounding environment or use no prior information. However, visual features extraction for positioning is not an easy task and requires a lot of computational resources. Therefore simpler and cheaper sound source localization has long been applied to areas such as speaker separation or airplane tracking using a variety of localization techniques. Sound source localization is a focus of this thesis.

4.1.1 Overview of sound localization approaches

Existing sound source localization procedures are based on either beamforming or time-difference of arrival (TDOA). Beamforming refers to any situation where the location estimate is derived directly from a filtered, weighted and summed version of the signal data received at the sensors. TDOA estimates the time delay between microphones receiving a signal, by comparing the signal properties using cross-correlation. Where beamforming combines any number of source signals in order to focus on sources in a chosen direction, TDOA compares the phase difference of two signals to detect a dominant source in any direction. Beamforming is a highly accurate method to detect sound sources in a small area. However, it requires a number of microphones and sophisticated signal processing. TDOA only requires two microphones and efficient signal processing to detect sources in a large range of locations. There are many different approaches within these general classes of localization procedures, each being developed with unique priorities to solve different problems.

4.1.2 Choice of localization method

Localization techniques generally improve with an increase in the number of microphones in the array, sometimes leading to large array systems. The benefits are especially true when adverse acoustic effects are present [175]. However, when acoustic conditions are favorable and the microphones are positioned judiciously, source localization can be performed adequately using a modest number of microphones. Performance is clearly affected by the array geometry, which is in turn dependent on the specific application conditions, hardware available and cost criteria. Passive localization systems are frequently TDOA-based, predominately due to their computational practicality and reasonable performance. Steered-beamformer strategies are computationally more intensive. In addition, the choice of the appropriate localization method is heavily influenced by signal properties such as: bandwidth (narrow or wideband signals), degree of correlation between signal components (coherent or incoherent) and the existence of a retardation effect.

A signal is classified as *narrowband* if the bandwidth is small compared to the inverse of the transit time of a wave front across the array. Otherwise a signal is called broadband (wideband). Traffic noise consists of a large frequency range (Chapter 3) and thus it is a broadband signal. Wideband or *broadband* problems can be decomposed into a set of narrowband ones by operating on the sensor data with a comb of narrowband filters. Krim and Viberk [97] provide an excellent review and comparison of many classical and advanced parametric narrowband localization techniques up to 1996.

Signal components arriving from different directions exhibit varying degrees of correlation ranging from totally uncorrelated or *incoherent* to fully correlated or *coherent* cases. In practical situations such as traffic noise or sonar, wave fronts show progressive loss of coherence with increasing spatial separation. This de-correlation results in an obscurity in the precise direction of arrival, i.e. the wavefront appears to arrive from a spread of angles centered around the true direction. The correlations between the sensors fall off as the separation between them increases. Such spatial de-correlation can result from propagation of the wave front through a refracting medium or from scattering. Paulraj and Kailath [147] investigated the sensitivity of the DOA estimates to spatial coherence or spatial de-correlation and proposed a solution to partially overcome this problem for narrowband signals, which could be extended to broadband signals. Coherent broadband direction of arrival was also examined by Abhayapala and Bhatta [7], where no preliminary knowledge of DOA angles, nor the number of sources to be estimated, were required.

Depending on the speed of the target relative to the speed of sound in air, the vehicle may have moved to a completely different position by the time its emitted acoustic signal arrives at the sensor array. In such a case, every observation of the vehicle location represents an estimate of the vehicle location history, rather than the current time. This so-called *retardation effect* complicates a solution to the problem of acoustic tracking of a maneuvering target from spatially distributed sensors. Dommermuth and Schiller [53] describe a maximum-likelihood (ML) technique to estimate the complete set of target motion parameters using an orthogonal array consisting of four microphones. Early work in DOA estimations included the early version of maximum-likelihood (ML) solution, but it did not become popular due to its high computational cost. A variety of techniques with reduced computations dominated the field. The more well-known techniques include the minimum variance method of Capon [32], the multiple signal classification (MUSIC) method of Schmidt [173] and the minimum norm of Reddi [162]. Lo and Ferguson [110] described a nonlinear least-squares method to estimate the complete set of target motion parameters that can be applied with an arbitrary sensor array.

For the purpose of traffic monitoring, vehicle tracking is not necessarily the primary objective. Once an individual vehicle is detected, provided it is distinguishable in some manner from other vehicles and its parameters are extracted, further tracking of the vehicle is superfluous to the purpose of the system. Tracking may be relevant in military applications, but for this work the disadvantages of the retardation effect are negligible. Therefore only bandwidth and coherence will be considered as relevant signal properties from this point onwards.

4.2 Beamforming

Passive sound detection and tracking has been a topic of research since World War II. The first approach in passive sound detection was space-time processing of data sampled at an array of sensors, called spatial filtering or beamforming. *Beamforming* is the name given to a wide variety of array processing algorithms that by some means focus the array's signal-capturing abilities in a particular direction [97]. It can be employed to separate signals according to their directions of propagation and their frequency content. Many research areas use beamforming in a variety of applications for the radiation or reception of energy, as summarized in Table 4.2. Due to its versatility and maturity, there is a vast array of publications on beamforming. Many tutorial papers [189, 37], books [29, 85, 188] and research papers [10, 136] have dealt with beamforming and localization. A brief overview of common beamforming methods is given in the remainder of this section.

4.2.1 Delay and sum beamforming

Delay-and-sum beamforming is the oldest and simplest array signal processing algorithm, often referred to as a conventional beamformer [85]. If a propagating signal is present, then the combined microphone outputs reinforce the signal by delaying the inputs by appropriate amounts and adding the inputs together to form a single output signal. Figure 4.1 shows a delay-and-sum beamformer linear combination of array sensor outputs. The output from a delay-and-sum beamformer may be described mathematically as

$$y(k) = \sum_{i=1}^{N} w_i x_i(k - \Delta_i), \tag{4.1}$$

where x is one of the sensor array outputs delayed by time Δ_i, w is the weighting and $y(k)$ is the combined signal for N array sensors. The weights determine the spatial filtering characteristics of the beamformer. They also separate signals with overlapping frequency content if they originate from different locations. The delays

Table 4.2: Beamformer applications [189]

Application	Description
RADAR	phased-array radar; synthetic aperture radar
Acoustics and SONAR	source localization and classification
Communications	directional transmission and reception sector broadcast in satellite communications
Imaging	ultrasonic; optical; tomographic
Geophysical exploration	earth crust mapping; oil exploration
Astrophysical exploration	high resolution imaging of the universe
Biomedical	tissue hyperthermia; hearing aids; fetal heart monitoring

that reinforce the signal are directly related to the length of time it takes for the signal to propagate between sensors, indicating the location of the sound source.

4.2.2 Filter and sum beamforming

More than one signal may be present in the wavefield measured by the sensors and noise can disturb the observations. To help remove these unwanted disturbances, additional linear filtering may be added to focus the array. The combination of these outputs is known as filter-and-sum beamforming, where the receiver weighting function depends on frequency. For each sensor in the array, the output is filtered

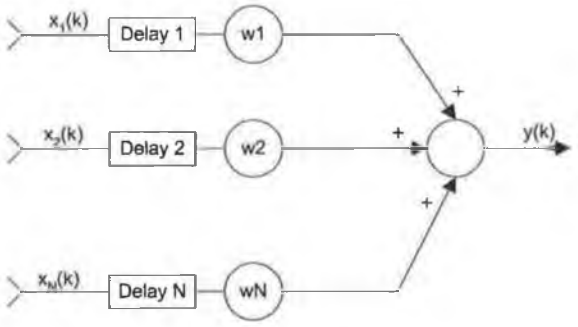


Figure 4.1: Delay-and-sum beamformer linear combination of array sensor outputs

with a weighting function wi(t) to yield a filtered signal. Then a delay-and-sum operation is performed on the filtered signal.

4.2.3 Frequency domain beamforming

In most beamforming applications, two assumptions simplify the analysis:

1. the signals incident on the array are narrowband;
2. the signal sources are located far enough away from the array so that the wavefronts impinging on the array can be modeled as plane waves (far-field assumption).

For many microphone array applications, the farfield assumption is valid, but not the narrowband assumption. An important dimension in measuring array performance is its size in terms of operating wavelength. Thus for high frequency signals a fixed array will appear large and the main beam will be narrow. However, for low frequencies the same physical array appears small and the main beam will widen. To overcome this problem, a beamformer must be used that is designed specifically for broadband applications. Typically broadband beamformers are implemented with a narrowband decomposition structure. The narrowband decomposition is often performed by taking a discrete Fourier transform of the data in each sensor channel using an FFT algorithm. The data across the array at each frequency of interest are processed by their own beamformer and inverse transformed back to the time domain. This is often termed frequency domain beamforming, where calculations are performed in the frequency domain. The derivation of the filters is what distinguishes beamforming methods.

4.2.4 Constant directivity beamformers

A specific class of broadband beamformers, called constant directivity beamformers (CDB), are designed such that the spatial response is the same over a wide frequency band. There have been several techniques proposed to design a CDB. Most techniques are based on the idea that at different frequencies, a different array should be used that has total size and inter-sensor spacing appropriate for that particular frequency.

4.3 Time delay of arrival localization

The signals received by microphones in an array due to an emitted sound are generally time-shifted versions of one another. The difference in time depends on the relative locations of the source and receivers as well as sound propagation speed. Additionally, there can be a variation in measured intensity level at different microphones. One of the earliest time-delay estimation approaches is based on obtaining the maximum cross-correlation between two microphone signals. Using this peak to estimate the time delay, together with knowledge of the microphone/source geometry, the source direction could be determined and in certain cases the location [33]. Figure 4.2 illustrates the process with two microphones receiving a time-delayed version of the same source signal. The angle of arrival is related to the time delay, which can be determined from the peak location in the cross-correlation sequence.

To improve the accuracy of localization, additional microphones may be used. When more than two microphones are used, the traditional TDOA approach involves two steps: a) compute TDOA for pairs of spatially separated microphones, b) combine these estimates in some manner to obtain the final source solution [145, 77, 172, 106]. There is a wealth of literature describing TDOA approaches applied to many different situations; near/far-field, indoor/outdoor, single/multiple sources, narrowband/wideband signals as well as for multiple microphone pairs. In the view of this thesis objective, only an outdoor TDOA source localization approach using two microphones is considered further.

4.3.1 Computing TDOA estimates

A landmark paper by Knapp and Carter in 1976 [93] described a *Generalized Cross Correlation* (GCC) time-delay estimation function that was central to future TDOA research. It assumes that the signals are uncorrelated, stationary Gaussian signals with no multi-path propagation and that noise sources have known statistics. It exploits the relationship between time-domain cross correlation and frequency-domain cross power spectral density function via a Fourier transform.

In the GCC time delay estimation function, the two signals to be cross-correlated are first transformed to the frequency domain and the cross power spectral density is obtained, before an inverse Fourier transform returns to the time domain.

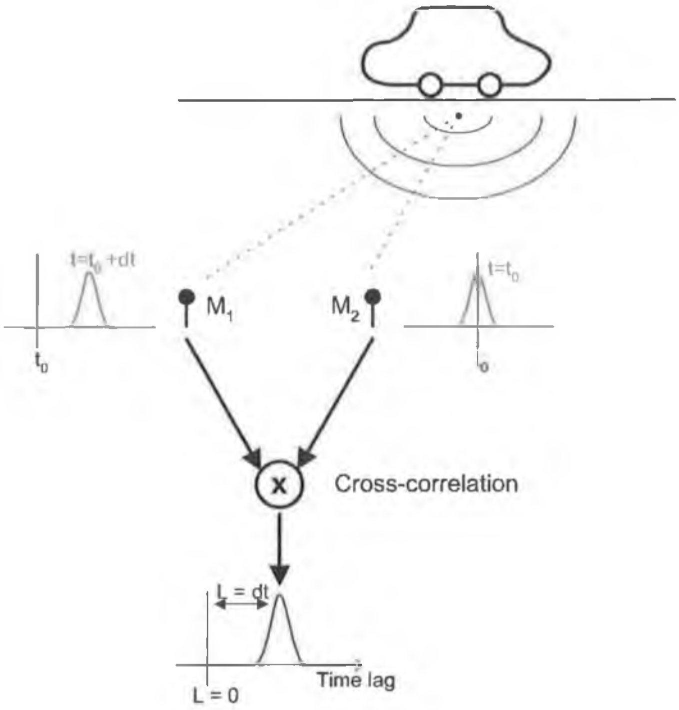


Figure 4.2: Illustration of the time-delay of arrival cross-correlation localization technique

Once reverberations rise above minimal levels, the simple GCC method is described as exhibiting dramatic performance degradations and becomes unreliable. Therefore the GCC method is modified in order to deal with distortions and to make the GCC function more robust. Knapp and Carter [93] describe a phase transform (PHAT) weighting. It effectively flattens the frequency domain cross-power spectral density magnitude - details are given in Section 5.3.3. If the noise spectrum of the received signal is known, maximum likelihood (ML) weights could be applied. However, detailed prior knowledge of the noise spectrum is generally not available.

The PHAT-weighting has received considerable attention as the basis of speech source localization systems [142, 145, 190], since the noise spectrum information is not required for its application. By placing equal emphasis on each component of the cross-spectrum phase, the resulting peak in the GCC-PHAT function corresponds to the dominant delay in the reverberated signal. It has the effect of eliminating the spectral magnitudes, resulting in a function entirely dependent on the phase of the cross-spectrum. Although the magnitude is less pronounced, the temporal resolution is much higher. While effective at reducing some of the degradations due to multi-path propagation, the PHAT method also accentuates components of the spectrum with poor signal-to-noise (SNR) ratio. Although the resulting cross-correlation functions often do have local maxima at the true time delay, they are not always global maxima, and can lead to erroneous time delay estimations. This leads to one or more of the time delay estimates for a microphone array being inaccurate and detrimentally affects the second step in the localization procedure. GCC is appealing for its simplicity and ease of implementation. However, it assumes a single-source model which limits its utilization to the multiple-source, reverberant environment problem. The GCC-PHAT method of time-delay estimation for source localization is described in further detail in Chapter 5.

4.3.2 Determining source location from TDOA estimates

Correctly determining a sound source location based on time-delay-of-arrival information requires more than simply calculating an appropriate cross-correlation sequence. The presence of a peak in the cross-correlation sequence simply indicates a strong inter-signal correlation. The location of a peak in a cross-correlation sequence may correspond to the time delay between two microphones receiving a similar signal from a single sound source, as desired. However, there may be a series of peaks in the vector, only one of which is the desired time delay estimate. A strong measure of confidence may be based on whether a peak under investigation continues to behave as expected over successive cross-correlation sequences. Selecting the largest peak in a cross-correlation sequence may not result in a correct time delay estimate. Equally, there is a disadvantage to making a premature decision on the time delay value that could result in useful information being discarded. The cross-correlation maximum is typically retained, however a secondary peak or even the entire cross-correlation sequence may also be relevant, especially if the global maximum is not the value of interest.

Due to the problems of multiple peaks, Bechler [18] considered the second peak in the GCC function for multi-source TDOA estimation with a microphone array. The principle of *least commitment* was used by Birchfield [25, 26] to preserve and propagate all the intermediate information to the end and make an informed decision at the very last step. This is similar to the novel use of an array of cross-correlation sequences in the shape-matching pattern extraction method developed in Chapter 7. Birchfield [26] did not make the plane-wave assumption that was inherent in his previous paper [25], making it much simpler and applicable to compact and non-compact microphone arrays. Griebel and Brandstein [66] maximise the entire GCC function over a set of potential delay combinations consistent with candidate source locations. The result was a procedure that combined the advantages offered by the PHAT weighting and a more robust localization procedure without dramatically increasing computational load. These may be viewed as a special case of the SRP-PHAT algorithm described by DiBiase et al. [51]. A less general technique was presented by Nishiura [133] which assumed that all microphone pairs are centered around the same location. Novel techniques to estimate source location are described in Chapter 7.

4.4 Examples of traffic monitoring systems

Examples of beamforming-based traffic monitoring systems are SmartSonic and SAS-1, previously introduced in Section 2.1.9. In this section, the implementation details are further discussed.

4.4.1 Beamforming-based traffic monitoring systems

SmartSonic and SAS-1 use a two-dimensional array of microphones and beamforming localization approach. The detection zone depends on the aperture size, frequency band and array geometry. The SmartSonic is tuned to 9kHz with a 2kHz bandwidth, ideally mounted between 10 to 30^o from the lowest point with a detection range of 6 to 11m. The SAS-1 sensor forms multiple detection zones with a microphone array and signal processing, to monitor up to 7 lanes when over the road or 5 at the roadside. Every 8ms the detection zones are checked and can be adjusted to 1.8m or 3.6m at a mounting height of 6-12m with the frequency range of 8-15kHz being processed.

The SAS-1 traffic monitoring system is an implementation of US Patent Number 5,798,983 [99]. This describes a multi-lane traffic monitoring system to measure vehicle presence, passage, speed and type using a 2-D microphone array. A conventional summing-line array beamformer is used for each row of sensors in the array. In order to satisfy the narrowband criteria, the signal is broken up into smaller frequency cells with adaptive complex weights applied to the resultant signal from each

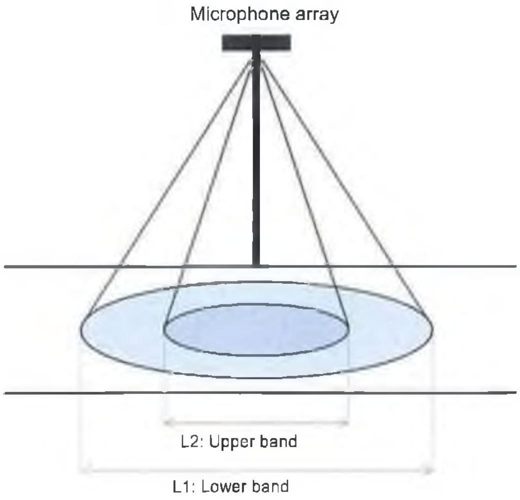


Figure 4.3: Beamforming approach to directionally monitoring road areas; from patent [99]

frequency cell before coherent summation. Lane positions are automatically determined by identifying the peaks (active lanes) and valleys (late separators/shoulders) of the averaged beam power response. Once the lane position is known, appropriate adaptive complex weights are applied to create a directional signal corresponding to a zone on each identified highway lane. The vehicle detection zone can be split into two areas corresponding to a lower and upper frequency band. The magnitude squared of specified signal frequency components within each band are summed to form the lower band and upper band adaptive power respectively. Since sensor directivity increases with signal frequency, the upper band detection zone is inside the lower band detection zone, as illustrated in Figure 4.3. Vehicle detection is performed by checking if lower and upper band signal magnitudes exceed certain thresholds. Speed is estimated from the time difference between the initial detection in the lower and upper bands as well as the difference in zone periphery location. A detailed version of the upper band signal can discriminate between discrete axle sound sources, enabling the measurement of vehicle length. This information is further used for vehicle classification according to length and axle position.

US Patent number 6,021,364 by Lucent Technologies Inc. [22] describes a highway vehicle presence detector where a binary signal is emitted during the presence of a vehicle. An array of microphones are arranged in a geometric arrangement[1] [69, 185]. In order to attenuate sounds emitted outside the desired detection zone, the micro-

[1] known as a Mill's Cross

phone array is shielded on all sides and from behind with a box-shaped mechanical baffle. The signals from the microphone array are combined using a classical beamforming technique designed to focus on a particular detection zone, similar to that described in patent [99]. The signal is bandpass filtered with a passband frequency bandwidth between 4 and 6kHz. If the magnitude exceeds a threshold, it is considered to represent the presence of a vehicle in the detection zone.

Both aforementioned patents make claims based on audio-based traffic monitoring systems using beamforming to perform source localization. Given the nature of patents, it is impossible to determine the effectiveness of such an approach on the patent applications alone. Although an insight into their methodology can be gained, no scientific evaluation or comparison against alternative methods is possible. Evaluations described in Section 2.2 included the use of an acoustic traffic monitoring product using beamforming, with acceptable traffic detection and speed estimation results. Based on these evaluations it can ascertained that such an approach produces useful results, although no publications have been found to date that further detail their performance for comparison.

4.4.2 TDOA-based traffic monitoring systems

There are no known audio TDOA-based traffic monitoring systems currently available, although there exists a number of publications describing how such a system could be implemented.

Couvreur and Bresler [45] used Doppler-based motion estimation for wide-band sources from single passive sensor measurements. Since only a single sensor was used, the approach involved the analysis of the acoustic signature to determine source speed and position. Poor performance is reported due to background noise, inappropriate stationarity point source assumption and inadequate modelling of sound propagation effects. These results mirror the results of signal feature classification experiments carried out by the author and described in Appendix A. Forren and Jaarsma [62] described cross-correlating the noise measured from vehicle tyres with three spatially separated roadside microphones, in order to track road vehicles. The possibility of measuring vehicle velocity and axle counting was described, based on observation of the measured cross-correlation matrix. However parameter extraction was performed by human observation with no automated pattern extraction, therefore results were limited. The approach, however, is roughly equivalent to the cross-correlation ground truth algorithm used by the author to test automated pattern extraction processes and described in Section 8.3.

Chen et al. [39] described a correlation-based traffic monitoring system using a large panel of microphones placed above a two-lane road. Many cross-correlation sequences were obtained from the microphones using the simple GCC method to distinguish sources in two directions; from lane to lane and along different road positions within each lane. Given the array size, this system may be more suitably applied in combination with a beamforming approach. They also mentioned how the trajectory steepness or slope ($\frac{d\tau}{dt}$) across the cross-correlation matrix is proportional to the velocity of the sound source and could be used to determine vehicle velocity. Additionally, it was mentioned how distinct sound traces observed from individual axles could enable axle separation. Similarly to Forren and Jaarsma [62], no automatic pattern extraction method was developed to extract the traffic indicators from the cross-correlation matrices. Therefore manual observation of the data would still be required to determine vehicle behaviour.

López-Valcarce and Pérez-Gonzáles [150] focused on determining vehicle velocity, based on a known road geometry and cross-correlation sequence from two microphones. From Equation 4.2 they noted that the velocity v could be estimated from the slope of $\Delta\tau$ at the closest point of arrival (CPA):

$$\frac{\partial \Delta\tau}{\partial t}|_{t=0} = -\frac{m}{Dc}v, \tag{4.2}$$

where v is vehicle velocity, m is the inter-microphone distance, D is the distance to the road, c is the speed of sound. However this equation is highly sensitive to errors in the determination of the slope since $\frac{m}{Dc}$ is usually very small. Therefore the velocity was obtained from the maximum likelihood estimate, similar to the approach of Betz [23] and Hassab et al. [72]. A full evaluation of the cross-correlation sequence is required for each candidate velocity, reducing the method efficiency. Vehicle movement during the propagation of its acoustic signature to the sensors must be taken into account, otherwise the estimate is biased for fast speeds and/or high-frequency components of the acoustic source. Therefore López-Valcarce et al. derived a delay error term [112]. López-Valcarce described tests based on real traffic and parameter evaluation [113]. The problem of CPA uncertainty was mentioned, since the location of the CPA has to be estimated. Additionally, the appropriate time window and sampling frequency must be determined as a trade-off between complexity and performance.

In conclusion, a limited number of publications have discussed and verified the capability of using cross-correlation data from microphone pairs to determine traffic parameters. However completed work did not include pattern extraction techniques for a fully automatic parameter extraction.

4.5 Comparison between localization methods

Localization approaches are developed with a specific purpose in mind, each with differing priorities depending on the envisaged application. Thus the choice of localization approach is bound within the constraints of the application. For all sound source localization approaches, the presence of multiple sources, excessive ambient noise or moderate to high reverberation levels in the acoustic field reduce performance. The merits of the different approaches and influencing factors are now summarized.

Beamforming has high computational requirements due to the large quantity of sensor and signal processing necessary. This prohibits its use in the majority of practical, real-time source locators. A further limitation is that the beamformer performance is directly dependent upon the size of the sensor array, where performance is suboptimal when using a small number of microphones. Sometimes it is not practical or possible to use an appropriately large array that would be required to obtain reasonable accuracy with a beamforming approach. Although the classical beamforming methods are useful to localize a narrowband source, the wideband nature of traffic noise demands a more complicated approach where the frequency band is either treated as a series of narrowband sources or a significant portion of the available frequency band is ignored. Steered-beamformer strategies are computationally more intensive than TDOA approaches, but tend to posses a robustness advantage and require a shorter time analysis interval. However, in real situations, the performance advantage of a steered beamformer is diminished because of incomplete knowledge of the signal and noise spectral content, as well as unrealistic stationarity assumptions.

The primary limitation of a TDOA cross-correlation approach is the reported inability to accommodate multi-source scenarios since these algorithms assume a single source model. However, Sturim et al [176] demonstrated that TDOA-based methods with short analysis intervals may be used to track several individuals in a conversational situation. Furthermore, cross-correlation experiments described in Chapter 8 demonstrate that multiple sound sources were successfully detected. As is described in Section 6.4, the detection area using source localization is limited to a small road length due to choice of parameter values. It is therefore impossible for two vehicles to occupy the same lane within the observed road length, minimizing the amount of multiple sources. Finally, the sharing of mutual information among dispersed sensor systems may resolve the multiple source issue.

An advantage of TDOA-based cross-correlation is the minimal adverse effect of weather, since wind and rain are spatially distributed sound sources and therefore produce very low peaks in the correlation domain. Primarily because of their computational practicality and reasonable performance, the bulk of passive talker localization systems in use today are TDOA-based. TDOA is better suited to vehicle tracking because it is more computationally efficient than beamforming whilst providing reasonably accurate performance. Since one core motivation is the economical advantage of a small number of microphones, the constraints of a microphone pair is the direction that was chosen. In that case localization methods such as beamforming are irrelevant since they require many more sensors for accuracy. For these reasons, it was decided to develop and investigate a TDOA-based localization approach as the basis for traffic monitoring.

Chapter 5

TDOA-based Source Localization Equations

Section 4.1 discussed various sound source localization techniques and the basic theory behind them. It included a TDOA approach, which was chosen as the most appropriate method to be implemented in this research project. This chapter describes the implementation of a TDOA cross-correlation approach in greater detail.

The purpose of the localization technique is to determine vehicle source direction or angle relative to the microphone array. By tracking the change in source angle over time, the vehicle velocity and direction can be measured. Individual vehicles occupying separate locations can be distinguished. Interference between different vehicles as well as superfluous noise is taken into consideration to optimize the technique.

5.1 Measured acoustic signal

A correlation-based process to calculate the time delay between two audio signals is now developed. Consider two microphones placed a distance m apart and distance D from the centre of the traffic lane. Assume the vehicle emits a source signal $s(t)$. If propagation distortion is disregarded, the signals received at the two microphones, $x_1(t)$ and $x_2(t)$ may be described as:

$$\begin{aligned} x_1(t) &= s(t) + n_1(t), \\ x_2(t) &= \alpha s(t+\tau) + n_2(t), \end{aligned} \tag{5.1}$$

where τ is the propagation delay between the two sensors and α is a relative attenuation constant (previously discussed in Section 3.2.1). $n_1(t)$ and $n_2(t)$ are assumed to represent zero mean Gaussian noise uncorrelated with $s(t)$. The characteristics of $x_1(t)$ and $x_2(t)$ directly influence the manner in which they may be compared to obtain the time delay. Therefore, the relevant properties of the acoustic signals recording road traffic data are now defined.

5.1.1 Measured signal properties

Data is considered to be *random* when future data values cannot be predicted within reasonable experimental error [20]. Aside from the unpredictability of traffic quantity, type and acoustical properties, there are a wide variety of factors affecting outdoor acoustics that cannot be controlled, some of which were described in Chapter 3. The measured audio signals of traffic are therefore described as random, since their future values cannot be anticipated.

When the statistical parameters of the data set change over time, a data is said to be *non-stationary*. In contrast, *stationary* signals are constant in their statistical parameters over time [20]. *Wide-sense stationary* processes have the looser requirement that the mean of the probability distribution and variance do not vary with respect to time. The acoustic signals from moving traffic are not stationary, since their statistical parameters certainly change over time. Quasi-stationary can be imposed by selecting a sufficiently short subsection of the signal, which then can be treated as if it were stationary for the purpose of analysis. By windowing the signal, a finite sequence with the desired length of data can be extracted from the signal to impose stationarity on a non-stationary audio signal. For example, speech has properties that are generally considered stationary for 20 to 30 ms and any expected frequency components are adequately resolved with this length [158]. Many existing signal processing methods such as Fourier transforms and cross-correlation assume at least wide-sense stationary. Existing analysis procedures for non-stationary data are substantially more limited, therefore it is beneficial to impose a quasi-stationarity assumption with the correct window size. In summary, the measured acoustic data is a random non-stationary discrete signal that can be assumed to be wide sense stationary due to appropriate windowing. The choice of window size is described in Section 6.2.4.

5.2 Cross-correlation

The cross-correlation for a particular time lag τ may be computed between the microphone pair as:

$$r_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} x_1(t)x_2(t-\tau)dt. \tag{5.2}$$

In practice, x_1 and x_2 are discrete sets of samples captured at a particular sampling frequency and not continuous functions such as $x_1(t)$ and $x_2(t)$ in Equation 5.2. The cross-correlation of discrete signals $x_1[n]$ and $x_2[n]$ is a sequence $r_{x_1x_2}[k]$, defined as [156, 143, 20]:

$$r_{x_1x_2}[k] = \sum_{n=-\infty}^{\infty} x_1(n)x_2(n-k) \text{ for } k = 0, \pm1, \pm2, ... \tag{5.3}$$

The index k is the time shift or lag parameter of the cross-correlation sequence $r_{12}(k)$, where k is a sampled version of time delay τ at a particular sampling frequency. The order of the subscripts in $r_{x_1x_2}$ indicate the direction in which one sequence is shifted relative to the other.

If one or both of the signals involved in the cross-correlation are scaled, the shape of the cross-correlation sequence does not change, instead the amplitudes of the cross-correlation sequence are scaled accordingly. Since scaling is unimportant, it is possible to normalize the cross-correlation sequence to the range from -1 to 1 so the sequence is independent of signal scaling. The normalized cross-correlation sequence is defined as:

$$\rho_{x_1x_2}[k] = \frac{r_{x_1x_2}(k)}{\sqrt{r_{x_1x_1}(0)r_{x_2x_2}(0)}}, \tag{5.4}$$

where $r_{x_1x_2}[k]$, or the normalized equivalent $\rho_{x_1x_2}[k]$, present the desired set of cross-correlation values between signals x_1 and x_2. Figure 5.1 shows the normalized cross-correlation sequence $\rho_{x_1x_2}[k]$ of two microphone signals of duration 22ms at a particular time instance. The inter-signal time delay τ can be observed in the graph as being the location of the prominent peak at 0. This represents the cross-correlation time lag or delay between the two input signals. Since the two input signals are from a pair of co-located microphones, τ represents the inter-microphone time delay of the dominant sound source that appears in both microphone signals.

In the case of a moving source such as a vehicle, τ changes as the source passes the microphone array. For this reason, it is interesting to observe the change in cross-correlation sequence over time by grouping the cross-correlation sequences obtained

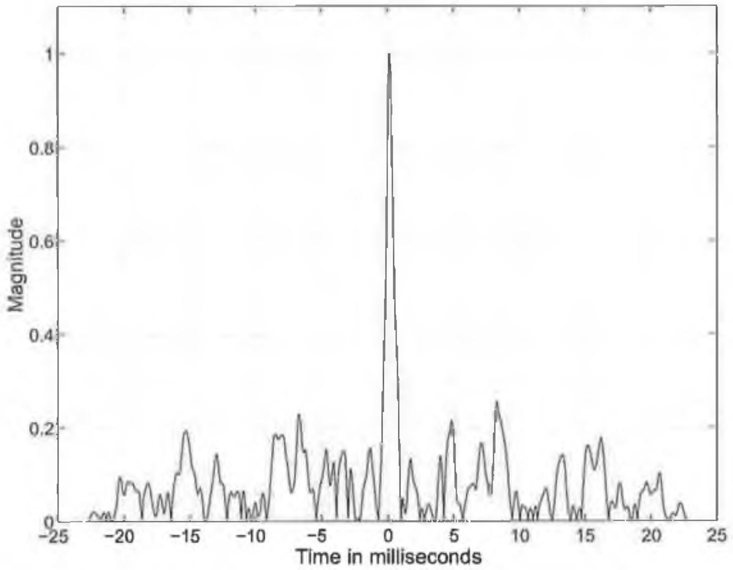


Figure 5.1: Normalized cross-correlation sequence for two microphone signals

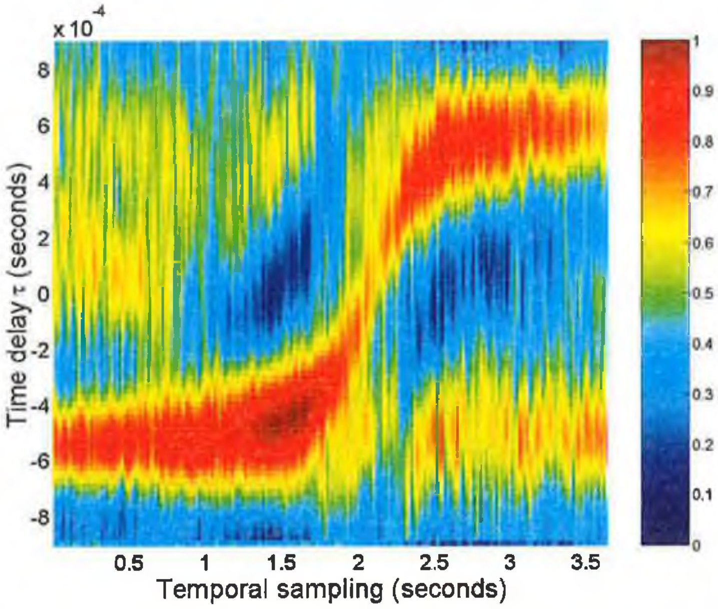


Figure 5.2: Cross-correlation array of a single vehicle

from different short-time windows into an array. Such a cross-correlation array is illustrated in Figure 5.2. The x-axis represents time passing and the y-axis is the cross-correlation sequence $r_{x_1x_2}[m]$ of two windowed microphone signals of length 22ms. The inter-microphone time delay τ can be observed as the cross-correlation peak value along the y-axis. Since the x-axis denotes progression in time, it can be observed that τ changes over time. As the vehicle approaches the microphone array, τ reduces in value until the vehicle is equidistant to the microphones. The vehicle continues in the same direction, moving away from the microphone array and hence causing τ to increase in value until it settles at the far-field maximum value.

The main difficulty with the classical time-domain cross-correlation approach is that the variance of the peak observed in Figure 5.2 is wide and at times it is difficult to determine the most accurate value of τ. Since the purpose is to accurately determine the time delay τ between two microphone signals, a technique is required that obtains a clearer and more defined value for τ in the cross-correlation sequence.

5.3 Frequency based cross-correlation

The cross-correlation sequence $r_{x_1x_2}[k]$ may also be obtained via the frequency domain, since the time-domain cross-correlation operation is related to the frequency-domain cross power spectral density function $G_{x_1x_2}(f)$ of the two signals x_1, x_2 [143, 156]. This is true for a wide-sense stationary signal, described in Section 5.1.1. Frequency-domain signals consist of magnitude and phase components. Since τ is represented by the phase difference between the two signals, frequency-domain weighting can be utilized to improve the phase representation of τ in the cross-correlation sequence. For this reason, frequency-domain cross-correlation is investigated with a view to determining τ from the phase information.

A finite-duration signal $x(n)$ of length $L \leq N$ in the time domain can be uniquely described as a set of N spectral samples in the frequency domain, where the linear discrete Fourier transform describes the mathematical relationships between these versions of the same signal.

$$X_1 = DFT[x_1(n)] = \sum_{n=0}^{N-1} x_1(n)e^{-j2\pi nk/N} \quad k = 0, 1, ..., N-1, \tag{5.5}$$

$$X_2 = DFT[x_2(n)] = \sum_{n=0}^{N-1} x_2(n)e^{-j2k\pi nk/N} \quad k = 0, 1, ..., N-1. \tag{5.6}$$

$X_1(k)$, $X_2(k)$ are Fourier transforms of $x_1(n)$ and $x_2(n)$, DFT represents the discrete Fourier transform of the time-domain signal. The auto power spectra $G_{x_1x_1}(f)$, $G_{x_2x_2}(f)$ and the cross power spectrum $G_{x_1x_2}(f)$ may be computed from $X_1(f)$, $X_2(f)$ as:

$$\begin{aligned} G_{x_1x_1}(f) &= X_1^*(f)X_1(f), & (5.7)\\ G_{x_2x_2}(f) &= X_2^*(f)X_2(f), & (5.8)\\ G_{x_1x_2}(f) &= X_1^*(f)X_2(f), & (5.9) \end{aligned}$$

where the asterisk indicates a complex conjugate. The time-domain cross-correlation $r_{x_1x_2}[k]$ between $x_1(n)$ and $x_2(n)$ is the equivalent of the frequency-domain cross power spectral density $G_{x_1x_2}(f)$.

$$r_{x_1x_2}[k] = x_1(n) \otimes x_2(n-k) \quad \leftrightarrow \quad G_{x_1x_2}(f) = X_1(f)^* \cdot X_2(f), \tag{5.10}$$

where $\otimes$ denotes convolution and $\cdot$ denotes multiplication.

The cross-correlation between signals $x_1(n)$ and $x_2(n)$ can therefore also be obtained by first transforming both signals to the frequency domain, multiplying the frequency-domain representations, and transforming the result back to the time domain. The transforms are most efficiently performed by using an FFT algorithm [143, 156]. If the number of terms in the sequences is sufficiently large, it is faster to use the frequency-based cross-correlation method than to calculate the cross-correlation directly in the time-domain. Frequency-domain spectral density estimation is used to cross-correlate data during the course of this work.

5.3.1 Weighting function

A weighting factor may be applied during frequency-domain cross-correlation in order to emphasize different aspects of the signal such as phase. This permanently modifies the resulting cross-correlation factor. The following discussion presents the motivation behind the use of a weighting function.

5.3.2 Requirements for weighting function

The microphone signals $x_1(t)$ and $x_2(t)$ consists of signal $s(t)$ and noise components $n(t)$. If τ is the propagation delay between the two sensors and α is a relative attenuation constant, we are writing:

$$\begin{aligned} x_1(t) &= s(t) + n_1(t), &(5.11)\\ x_2(t) &= \alpha s(t-\tau) + n_2(t). &(5.12) \end{aligned}$$

The cross-correlation of $x_1(t)$ and $x_2(t)$ can be described as

$$r_{x_1x_2}(t) = \alpha r_{ss}(t-\tau) + r_{n_1n_2}(t). \quad (5.13)$$

The Fourier transform of Equation 5.13 gives the cross power spectrum

$$G_{x_1x_2}(f) = \alpha G_{ss}(f)e^{-j2\pi f\tau} + G_{n_1n_2}(f). \quad (5.14)$$

If $n_1(t)$ and $n_2(t)$ are uncorrelated, then $G_{n_1n_2}(f) = 0$. Since multiplication in one domain is a convolution in the transformed domain, it follows that the time-domain equivalent of Equation 5.14 is

$$r_{x_1x_2}(t) = \alpha r_{ss}(t) \otimes \delta(t-\tau). \quad (5.15)$$

Equation 5.15 consists of the desired time delay τ in the delta function, convolved with source signal auto-correlation $r_{ss}(t)$. $r_{ss}(t)$ has the detrimental effect of effectively spreading the delta function, thus broadening the time-delay peak of interest. The broadening effect of the source signal autocorrelation was also described by Knapp + Carter [93]. In addition to this problem, any slightly correlated noise further complicates the measurement of τ. For multiple time delays, one delta function can spread into another, thereby making it impossible to distinguish peaks or delay times. The spreading has the effect of broadening the true cross-correlation peak, an effect that should be avoided or at least minimized. A weighting function $\psi(f)$ is desired that improves the accuracy of the time delay estimate by reducing the auto-correlation spreading effects, and the effects of correlated noise. The generalized cross-correlation then becomes:

$$r_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} \psi(f)G_{x_1x_2}(f)e^{j2\pi f}df, \quad (5.16)$$

where $G_{x_1x_2}(f) = X_1(f)^*X_2(f)$. The chosen weighting function $\psi(f)$ should ensure a large sharp peak in $r_{x_1x_2}(\tau)$, in order to achieve good time-delay resolution. However, sharp peaks are more sensitive to errors introduced by finite observation time, particularly in cases of low S/N ratio. To minimise $r_{ss}(t)$, the source signal characteristics should be suppressed. It is impossible to specify $G_{ss}(f)$ or $r_{ss}(t)$, since no prior knowledge of the source signal characteristics is known. Therefore a general weighting factor is required that suppresses $r_{ss}(t)$, making $\delta(t-\tau)$ more defined [93].

5.3.3 Description of weighting function

The inter-microphone time delay τ is the desired value of interest and is purely a time difference. The magnitude of values in the cross-correlation sequence is not relevant except as a means to determine the strongest candidate for τ. However, the magnitude of the cross-correlation sequence includes contributions from the source signal auto-correlation $r_{ss}(t)$ as described in Section 5.3.2. In an attempt to minimize the effect of $r_{ss}(t)$, $G_{x_1x_2}(f)$ is divided by its magnitude component $|G_{x_1x_2}(f)|$. This is equivalent to setting the magnitude of $G_{x_1x_2}(f)$ to 1 while preserving the phase information. The modified signal is then transformed to the time domain. Equation 5.17 describes the modified cross-correlation sequence.

$$r_{x_1x_2}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{X_1(f)^*X_2(f)}{|X_1^*(f)X_2(f)|}e^{j2\pi f\tau}df = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{G_{x_1x_2}(f)}{|G_{x_1x_2}(f)|}e^{j2\pi f\tau}df. \tag{5.17}$$

The applied weighting function is therefore $\frac{1}{|G_{x_1x_2}(f)|}$. The weighting function can be considered as a pre-whitening filter applied to the cross-power spectrum in order to weight the magnitude value against its SNR. The weighting function chosen is equivalent to the Phase Transform (PHAT) weighting described by Knapp and Carter [93] and used in [170, 171, 98]. It requires no prior knowledge of the signal or noise characteristics and assumes signal stationarity.

Figure 5.2 shows the simple cross-correlation array resulting from a vehicle passing, while Figure 5.3 uses the same original data to obtain the weighted cross-correlation array. In both images it can be observed that a passing vehicle generates an S-shaped pattern or signature in the cross-correlation array. For the simple cross-correlation array it is difficult to define the exact source location or value of τ per time instance, as there are a range of possible values due to the large width of the cross-correlation peak. The pattern created by a passing vehicle in the weighted array is more defined and distinguishable from background noises, despite the magnitude being lower. The maximum of the cross-correlation peak in the simple cross-correlation array is typically much greater than the maximum of the weighted cross-correlation peak, relative to their respective means. This is due to the flattening of the frequency-domain magnitude.

As the single sound source approaches the microphone array in the weighted cross-correlation array, the pattern "splits" into two separate sources. Beyond a certain

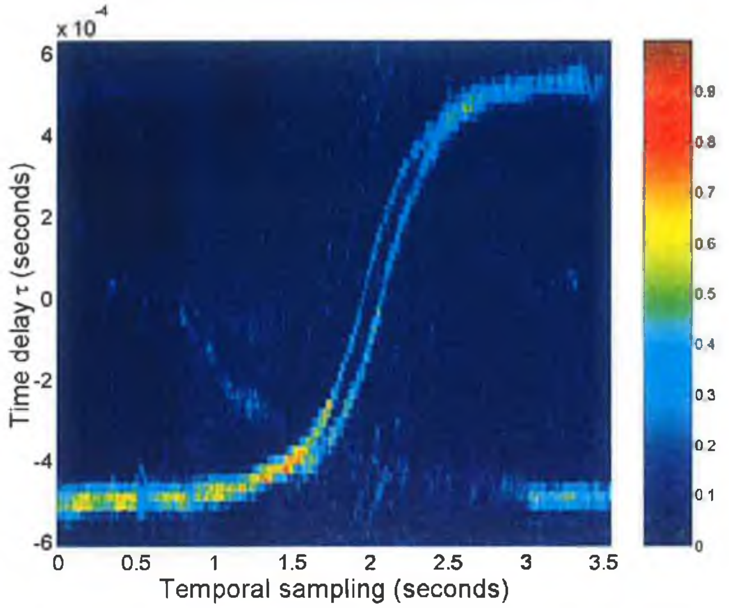


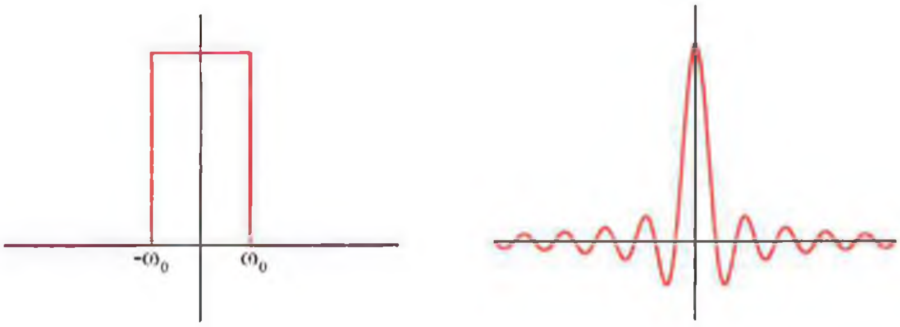
Figure 5.3: PHAT-weighted cross-correlation array of a single passing vehicle, visible as front and rear sound sources when in close proximity

proximity these sources merge once more into a single source. The sound generated by individual axles can be distinguished as two sources, provided the vehicle is within a certain angular range to the microphone array. This fidelity is not visible in the simple cross-correlation array.

The suppression of the frequency spectral magnitude effectively places correlated noise components in the received signal cross-power spectral phase on an equal level with the source signal. For this reason the PHAT weighting is often described in publications as being particularly suited to scenarios with low noise levels. Experimental results in Section 8.3 confirm that vehicles were successfully detected with a high accuracy when using the described weighting, despite significant background noise. Due to the high rate of correct vehicle detection and precise value of measured τ, it was found that the proposed weighting is suitable for audio-based traffic monitoring. Therefore, the weighted cross-correlation array is used to determine time delay τ.

Side-effect of weighting function

A relevant side effect of applying the weighting to the frequency-domain cross-spectral density is that the flattened magnitude component then approximates a pulse DC characteristic. A pulse DC frequency-domain characteristic transforms to the time domain as a sinc function overlaid on the phase difference information containing the inter-microphone time delay. This effect was not described in research publications describing the use of the PHAT weighting function. Therefore, although the weighting has the effect of minimizing the time-delay peak spreading, it introduces a sinc function that detracts from the time-delay peak, particularly in the region of $\tau = 0$. The magnitude and shape of the sinc function is directly related to the frequency-domain signal magnitude.



$F(\omega) = \prod$ where $\omega = \pi f_s$ $f(t) = sinc(\omega_0) = \frac{sin[\pi f_s t]}{\pi f_s t}$

Figure 5.4: (a) rectangle (b) sinc function

By applying weighting, the magnitude of the frequency domain cross-power spectral density estimate is effectively flattened. The frequency-domain magnitude is not an infinitely-long signal, but rather a finite sequence limited by the window size. The inverse Fourier transform of a finite sequence of constant magnitude (i.e. a rectangular signal) is an infinite sinc function, illustrated in Figure 5.4. As the rectangular pulse becomes taller and narrower, the sinc function grows flatter and wider. The effect of the sinc function can be observed in Figure 5.5 as a peak at $\tau = 0$.

The irrelevant sinc function in the cross-correlation sequence detracts from the cross-correlation peak due to τ. However, the largest component of the sinc function is always in the region where $\tau = 0$. Furthermore, the magnitude of the overlayed sinc function is quantifiable from a known frequency-domain pulse DC response. With this knowledge, peaks around $\tau = 0$ can be ignored or the central magnitude adjusted to remove the strong effect of the sinc function at that location. The beneficial accuracy of the weighting still outweighs the disadvantages, resulting in a weighted cross-correlation sequence that contains sharp peaks representing the time

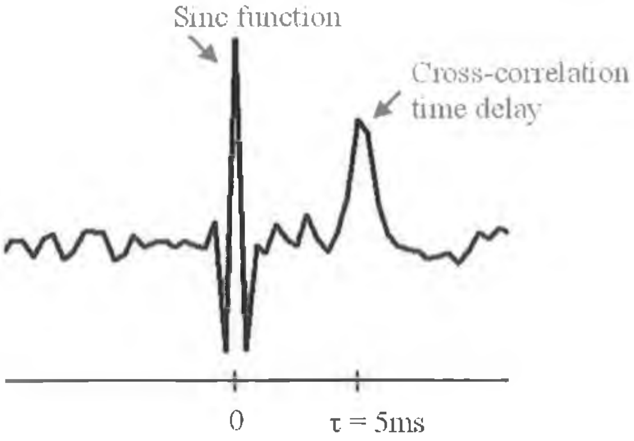


Figure 5.5: Illustration of the sinc function in a cross-correlation sequence

delay of source signals.

5.3.4 Interpolation of cross-correlation sequence

Each cross-correlation sequence consists of values in a series of time-delay bins. Since the audio signals are discretely sampled with a sampling frequency f_s, each time-delay bin is $1/f_s$ wide. Therefore cross-correlation values are only accurate to within half a bin. Interpolation can be used to decrease the bin size by interspersing calculated cross-correlation values with interpolated values. These approximate values may be obtained using a number of interpolation methods, all of which are determined by modelling the general behavior of surrounding samples. In this manner, the samples are padded with approximate values, thereby increasing the bin resolution.

Interpolation does not increase the accuracy of the original data samples, it only aids approximating cross-correlation estimation between samples. In this manner interpolation increases intra-sample resolution, where the interpolation factor is the ratio of the output rate to the input rate. Interpolation with a factor of 4 was applied to each cross-correlation sequence, resulting in each bin being subdivided into 4 segments. Cubic spline interpolation was used, since it chooses piecewise cubic polynomials between the data points to return a smoother estimate and incur a smaller error than linear interpolation [47, 75]. Figure 5.6 illustrates the cubic spline interpolation of data.

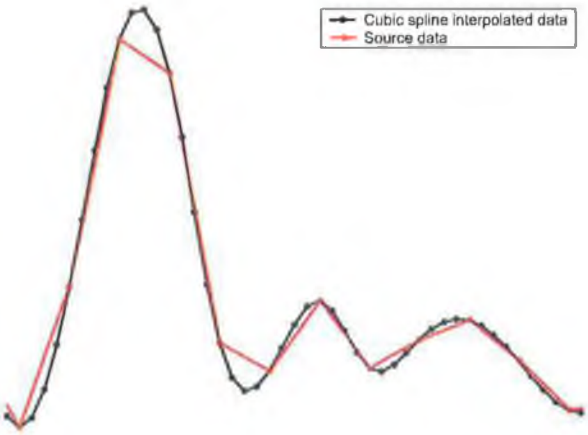


Figure 5.6: Cubic spline interpolation of data

5.4 Cross-correlation array characteristics

The formation of an interpolated, weighted data array has been described that contains pertinent information indicating the cross-correlation time delay τ between two microphone signals. When a dominant sound source is in close proximity to the microphones, its relative location can be determined based on the value of τ for that time instance. Algorithm 1 summarizes the steps involved in calculating the cross-correlation sequence for a particular time window. Successive cross-correlation sequences are then grouped to form an array of cross-correlation data.

Algorithm 1 Cross-correlation via the frequency domain with weighting and interpolation

```
1. Compute the N-point discrete Fourier transform to obtain $X_1[k]$ and $X_2[k]$
2. Compute $G_{x_1 x_2} = X_1^*[k]X_2[k]$, where $X_1^*$ is the complex conjugate of $X_1$
3. Multiply by the weighting function $\psi = 1/|G_{x_1 x_2}|$
4. Compute the inverse discrete Fourier transform of $\psi G_{x_1 x_2}$ to obtain $r_{x_1 x_2}$
5. Interpolate $r_{x_1 x_2}[m]$ to decrease the inter-sample distance
```

A description of the characteristics of the cross-correlation array is now given, so that an appropriate method can be selected and optimized to interpret the array correctly and determine the correct value of τ for each time instance. Consider Figure 5.7 containing a cross-correlation array of passing vehicles. It provides an illustration of the following observations regarding available data. Further images of cross-correlation array data are presented in Section B.1 of Appendix B.

The time-delay peak in the pattern created by a passing vehicle is not always distinguishable, as it can temporarily be hidden. If the change in peak location is being tracked over time relative to its previous location, this will cause the track to be intermittently or prematurely lost. Two separately detected patterns may be due to a single vehicle trace that is partially concealed in a noisy cross-correlation array. Furthermore, the sound generated by the front and rear of a vehicle is observed by the microphone array as one single overall sound when distant, but is resolved into two separate sources when in close proximity. In this manner, multiple sound sources may originate from a single vehicle, introducing the need to link related disparate data.

There is an outer limit to the location of peaks in the cross-correlation sequence due to vehicular noise. This limit is imposed by the distance between the two microphones and will be discussed in Section 6.2.1. Peaks at this limit indicate the presence of distant vehicles that are not in close enough proximity to the array to identify a unique position. Therefore, peaks at or beyond this outer limit do not contain relevant data describing the change in vehicle location, and are ignored for the purposes of vehicle monitoring.

Sometimes there are multiple time-delay peaks present in the cross-correlation sequence, each corresponding to different correlated sound sources. An expected time-delay peak in the cross-correlation sequence may have a smaller magnitude than another peak, due to two or more separate sound sources. This is shown in Figure 5.7 at approximately the 3 second mark, where an approaching sound source dominates the cross-correlation array and effectively "hides" the departing sound source.

Due to signal quantization, values jump abruptly. The chosen sampling frequency and interpolation rate dictate the level of quantization. Evidence of a moving source is often located in multiple successive bins of the discrete cross-correlation array for a given τ. This indicates that the measured data is not a smoothly undulating curve like a continuous function, but rather a jagged discrete series of steps with different lengths.

The significant noise around the 0-time delay (described in Section 5.3.3) detracts

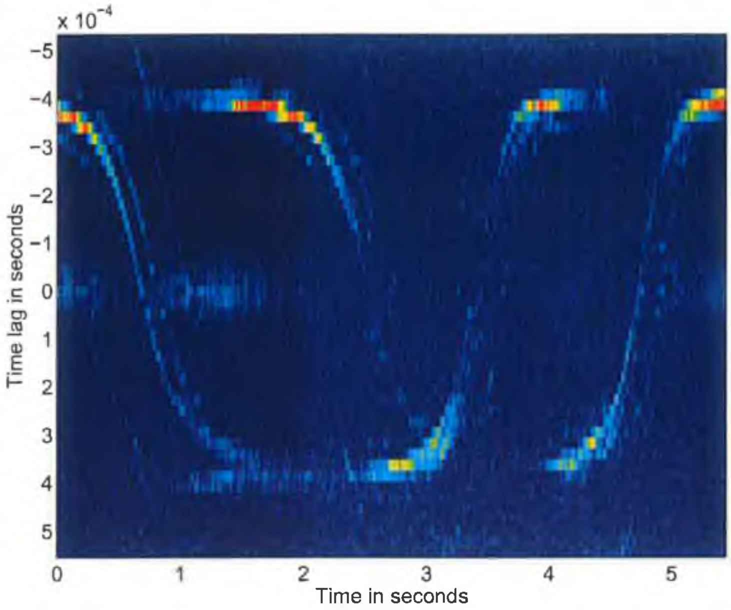


Figure 5.7: GCC-PHAT cross-correlation matrix of four vehicles

from more useful data at other locations. It is necessary to either compensate or ignore peaks in this region due to the sinc function. The absolute magnitude of the cross-correlation array is less relevant than the relative magnitude. In other words, when detecting a peak, its roll-off and nearest local maxima are much more informative than the absolute peak magnitude.

Although cross-correlation is a single-source model, it succeeds in presenting evidence of multiple sources in a single cross-correlation sequence. This is particularly apparent in Section B.1 of Appendix B where numerous images of simultaneous vehicles and airplanes are shown. In some cases, 3 distinct patterns are visible in the same region. To some degree, the strength of evidence is weakened when there are multiple sources, i.e. the protrusion of a particular time-delay cross-correlation peak is less distinct from the rest of the sequence if there are two or more such time-delay peaks. Nevertheless, it can be argued from the evidence that the cross-correlation array may be successfully utilized to monitor multiple sources for traffic monitoring. It is the task of the pattern extraction methods described in Chapter 7 to succeed in detecting these multiple sources in the cross-correlation array.

5.5 Conclusions

A method to obtain cross-correlation information reflecting the inter-signal time delay and therefore source location has been described in this chapter. Based on the GCC-PHAT algorithm, the described approach is applied to traffic signals obtained with a microphone pair.

Cross-correlation via the frequency domain is faster and allows the possibility of prioritising the phase information in the data. This is performed by flattening the frequency-domain cross-power spectral density magnitude with the weighting function. As a result, the spreading effect of the source signal autocorrelation function on the time-delay peak of interest is counteracted. By weighting and interpolating the data, the desired time-delay peak becomes sharper and more precise. Successive cross-correlation sequences combined in an array illustrate the movement of source location by the change in time delay through the array. The data contained in the cross-correlation array provides valuable information regarding moving sources and is more powerful than individual cross-correlation sequences. For this reason, the cross-correlation array is a primary source of information to localize traffic in the proposed automatic monitoring system. Chapter 7 describes how the cross-correlation array is analysed.

Chapter 6

Moving source geometrical modelling

Chapter 4 describes sound source localization techniques and Chapter 5 describes the use of localization to track the location of a moving sound source as it passes a stationary microphone array placed adjacent and parallel to the road. Based on known microphone array geometry, it is possible to model such a moving source. The benefits of modelling the sound source behavior include the ability to perform simulations for a range of variables and parameters such as source velocity. This gives greater insight into expected results and a better understanding of the scenario. Secondly, real data can be compared against an accurate model for verification, or to ascertain the parameter values for that particular case. Thirdly, the influence of parameter choice on accuracy of results can be estimated, reducing the need for exhaustive measurements and tests. For these reasons, a series of equations derived to model source location are described in this chapter.

Section 6.1 describes in detail how the equations are derived, culminating in a summary of the relevant equations that can be applied to the most general case in Section 6.1.6. Using the relevant equations, the accuracy and limitations of such a time-delay estimation approach are modelled in Section 6.2.

6.1 Derivation of source location equations

By cross-correlating signals in a microphone pair, the time-delay of arrival can be measured, as is described in Chapter 4. This inter-microphone time delay, τ, is directly related to the source direction or angle θ. As the source moves, the time delay changes accordingly. This relationship between τ and source location or angle

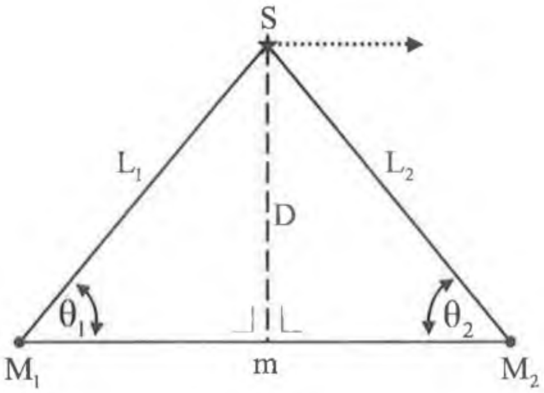


Figure 6.1: Reference point microphone array geometry where $L_1 = L_2$

θ can be derived firstly for a single reference time and location, then subsequently as a function of time when the source is moving through different locations at a constant velocity. The following section describes the reference case, where time t is zero and the source is directly opposite the microphone array.

6.1.1 Reference location equations

Consider the ideal case of two microphones picking up one sound source in Figure 6.1, where the reference point is when $\tau = 0$ i.e. the source is halfway between the microphones. In this case $L_1 = L_2$ and $\theta_1 = \theta_2 = \theta$. Microphones M_1 and M_2 receive the acoustical signal $r_{x_1}(t)$ and $r_{x_2}(t)$ at times t_1 and t_2 respectively. The delay in time between the microphones receiving the signal from source S can be described as

$$\tau = |t_2 - t_1|. \tag{6.1}$$

In the ideal situation shown in Figure 6.1, there is no delay in the microphones receiving the same signal, since $L_1 = L_2$, making $\tau = 0$. Time t is set at 0 for this reference point. The angle θ can be described as

$$\cot\theta = \frac{m/2}{D}, \tag{6.2}$$

$$\theta = \cot^{-1}\frac{m}{2D}. \tag{6.3}$$

$\cot\theta$ is used rather than $\tan\theta$ because $\tan\frac{\pi}{2}$ is not defined. Although $\cot 0$ and $\cot\pi$ are also not defined, the inter-microphone time delay τ is far away in such situations and therefore less interesting. As a result, the angle would never be sought for $\cot 0$ and $\cot \pi$. Equation 6.4 (relating to the case when the source is equidistant from microphones M_1 and M_2) can be used to calculate $\theta = \theta_1$ for $\tau = 0$ for a particular inter-microphone distance m. The next section considers the generalized case where $\theta_1 \neq \theta_2$.

6.1.2 Generalized triangle

Consider two microphones picking up one sound source as shown in Figure 6.2. In this case $L_1 \neq L_2$ and therefore $\theta_1 \neq \theta_2$. It is intended to derive an equation describing either θ_1 and/or θ_2 in terms of known parameters such as m or D, and parameters such as the location of S. From the left-hand and right-hand right-angle triangles respectively,

$$\begin{aligned} \gamma_1 &= 90^o - \theta_1, \\ \gamma_2 &= 90^0 - \theta_2. \end{aligned}$$

Since

$$\begin{aligned} \tan\gamma_1 &= \frac{m_1}{D}, \\ m_1 &= D\tan\gamma_1\theta_1, \end{aligned}$$

and

$$m_2 = D\tan\gamma_2\theta_2,$$

then

$$\begin{aligned} m &= m_1 + m_2 = D\tan\gamma_1 + D\tan\gamma_2, \\ m &= D\tan(90^o - \theta_1) + D\tan(90^o - \theta_2). \end{aligned} \tag{6.4}$$

Finally

$$m = D(\cot\theta_1 + \cot\theta_2). \tag{6.5}$$

Using Equation 6.5 relating m to angles θ_1 and θ_2, the next step is to describe the inter-microphone time delay τ as a function of θ_1 and θ_2. Since

$$\sin\theta_1 = \frac{D}{L_1} \Rightarrow L_1 = \frac{D}{\sin\theta_1} \quad \text{and} \quad \sin\theta_2 = \frac{D}{L_2} \Rightarrow L_2 = \frac{D}{\sin\theta_2},$$

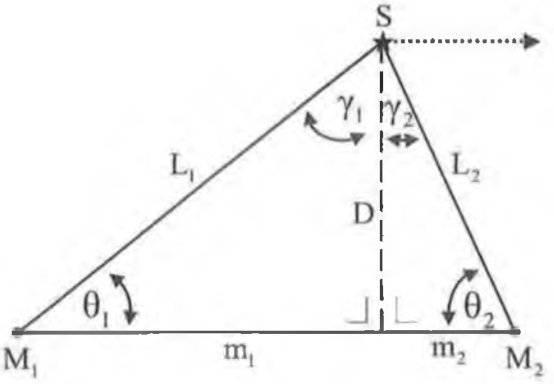


Figure 6.2: General microphone array geometry

then τ can be written as

$$\begin{aligned}
\tau &= \frac{1}{c}(L_1 - L_2), \\
\tau &= \frac{1}{c}\left[\frac{D}{\sin\theta_1} - \frac{D}{\sin\theta_2}\right], \\
\tau &= \frac{D}{c}\left[\frac{1}{\sin\theta_1} - \frac{1}{\sin\theta_2}\right].
\end{aligned} \tag{6.6}$$

Equation 6.6 describes τ in terms of the two angles θ_1 and θ_2. Therefore the next step is to replace either θ_1 or θ_2. To write τ in terms of θ_1, $1/\sin\theta_2$ is replaced with $\sqrt{1+\cot^2\theta_2}$ in Equation 6.6. Consider the following equation:

$$\sin^2\theta + \cos^2\theta = 1 \qquad (\theta \neq 0, 180^o). \tag{6.7}$$

Divide 6.7 across by $\sin^2\theta$

$$1 + \cot^2\theta = \frac{1}{\sin^2\theta}. \tag{6.8}$$

Then obtain the square root

$$\frac{1}{\sin\theta} = \pm\sqrt{1+\cot^2\theta}. \tag{6.9}$$

Since $0 < \theta < 180^0$, then $\sin\theta \geq 0\ \forall\theta$. Because of this we take the positive square root:

$$\frac{1}{\sin\theta} = \sqrt{1+\cot^2\theta}. \tag{6.10}$$

Therefore Equation 6.6 can be rewritten as follows:

$$\tau \;=\; \frac{D}{c}\left[\frac{1}{\sin\theta_1} - \sqrt{1+\cot^2\theta_2}\right].$$

Noting that $\cot\theta_2 = \frac{m}{D} - \cot\theta_1$ from Equation 6.5, θ_2 can be removed:

$$\tau = \frac{D}{c}\left[\frac{1}{sin\theta_1} - \sqrt{1+\left(\frac{m}{D} - \cot\theta_1\right)^2}\right]. \tag{6.11}$$

Similarly,

$$\tau = \frac{D}{c}\left[\sqrt{1+\left(\frac{m}{D} - \cot\theta_2\right)^2} - \frac{1}{\sin\theta_2}\right]. \tag{6.12}$$

Equations 6.11 and 6.12 describe the inter-microphone time delay τ in terms of either θ_1 or θ_2 respectively. Both equations are transcendent equations, meaning that θ can be solved numerically but not analytically. Given a known microphone array placed according to Figure 6.2, any time delay for every angle θ_1 or θ_2 can be calculated based on these two equations.

6.1.3 Moving source

The sound source S is not ordinarily stationary but rather passes the microphone array, moving from one location to the next. The goal of this section is to derive the change in source location as a function of time or distance to provide information about the source velocity and direction of travel. For this reason, how θ changes from S to S' with respect to distance d ($\Delta\theta_1$) is now considered, shown in Figure 6.3. $\Delta\theta_1$ is described as follows:

$$\Delta\theta_1 = \theta_1 - \theta_1^{'}. \tag{6.13}$$

Based on the assumption that the microphone array is parallel to the road, $\angle \mathrm{SS}'M_1 = \theta_1^{'}$ and $\angle M_2SS' = \theta_2^{'}$. Consider the general triangle shown in Figure 6.4(a). Now consider triangle $\Delta M_1SS'$ only, shown in Figure 6.4(b). The angle at S is

$$180^o - \theta_1^{'} - \Delta\theta_1 \;=\; 180^o - \theta_1^{'} - (\theta_1 - \theta_1^{'}) = 180^o - \theta_1.$$

Using the Projection Theorem $c = a\cos\beta + b\cos\alpha$, the following can be written:

$$d = L_1\cos(180^o - \theta_1) + L_1^{'}\cos\theta_1^{'}.$$

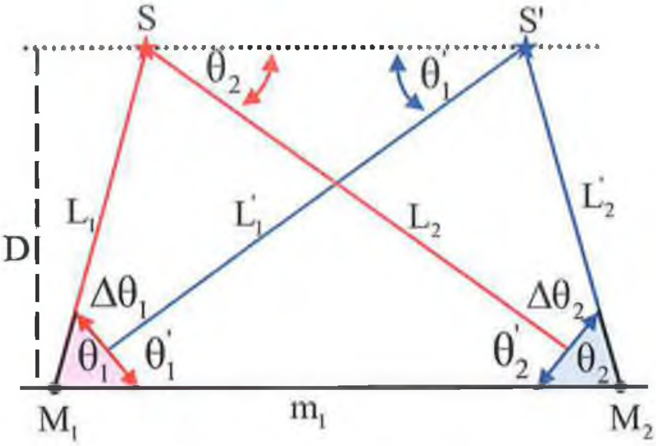


Figure 6.3: Road geometry with sound source at different locations

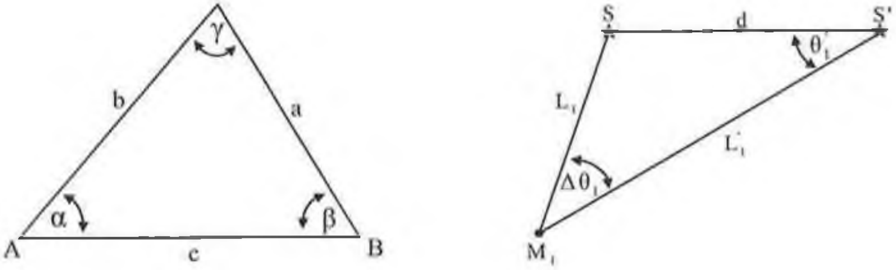


Figure 6.4: (a) Projection Theorem Triangle (b) Triangle ΔM1SS′

Since

$$L_1 = \frac{D}{sin\theta_n},$$

and

$$L'_1 = \frac{D}{sin\theta'_1},$$

then L_1 and L'_1 can be replaced, resulting in:

$$\begin{aligned} d &= \frac{D}{\sin\theta_1}\left[-\cos\theta_1\right] + \frac{D}{\sin\theta'_1}\left[\cos\theta'_1\right], \\ d &= D[-\cot\theta_1 + \cot\theta'_1]. \qquad (6.14) \end{aligned}$$

Equation 6.14 would be more useful if d was expressed only in terms of θ_1 and not θ'. To obtain such an equation it can be taken into account that $\theta'_1 = \theta_1 - \Delta\theta_1$, this substitution into Equation 6.14 results in:

$$d = D[\cot(\theta_1 - \Delta\theta) - \cot\theta_1]. \qquad (6.15)$$

Knowing that $\cot\alpha - \cot\beta = \frac{-\sin(\alpha-\beta)}{\sin\alpha\sin\beta}$ means that $\cot(\theta_1 - \Delta\theta_1)$ can be replaced in Equation 6.14, then

$$d = D\frac{-\sin(\theta_1 - \Delta\theta_1 - \theta_1)}{\sin(\theta_1 - \Delta\theta_1)\sin\theta_1} = D\frac{-\sin(-\Delta\theta_1)}{\sin(\theta_1 - \Delta\theta_1)\sin\theta_1}. \qquad (6.16)$$

Using the relationship $\sin(-x) = -\sin(x)$ gives

$$d = D\frac{\sin\Delta\theta_1}{\sin(\theta_1 - \Delta\theta_1)\sin\theta_1}. \qquad (6.17)$$

Since $\sin(\alpha - \beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta$ we get

$$d = D\frac{\sin\Delta\theta_1}{[\sin\theta_1\cos\Delta\theta_1 - \cos\theta_1\sin\Delta\theta_1]\sin\theta_1}. \qquad (6.18)$$

Dividing above and below the line by $\sin\Delta\theta_1$ we get

$$\begin{aligned} d &= D\frac{1}{\sin\theta_1[\sin\theta_1\frac{\cos\Delta\theta_1}{\sin\Delta\theta_1} - \cos\theta_1\frac{\sin\Delta\theta_1}{\sin\Delta\theta_1}]}, \\ &= D\frac{1}{\sin\theta_1[\sin\theta_1\cot\Delta\theta_1 - \cos\theta_1]}. \end{aligned}$$

Cross-multiply to get

$$\begin{aligned}\sin^2\theta_1\cot\Delta\theta_1 - \sin\theta_1\cos\theta_1 &= \frac{D}{d},\\ \sin^2\theta_1\cot\Delta\theta_1 &= \frac{D}{d} + \sin\theta_1\cos\theta_1.\end{aligned}$$

Then divide by $\sin^2\theta_1$ giving

$$\begin{aligned}\cot\Delta\theta_1 &= \frac{D}{d}\frac{1}{\sin^2\theta} + \frac{\cos\theta_1}{\sin\theta_1},\\ &= \frac{D}{d}\frac{1}{\sin^2\theta} + \cot\theta_1.\end{aligned}$$

Since $\frac{1}{\sin^2\theta_1} = 1 + \cot^2\theta_1$, replace $\frac{1}{\sin^2\theta}$ to get

$$\begin{aligned}\cot\Delta\theta_1 &= \frac{D}{d}(1+\cot^2\theta_1) + \cot\theta_1,\\ \Delta\theta_1 &= \cot^{-1}\left[\frac{D}{d}(1+\cot^2\theta_1) + \cot\theta_1\right].\end{aligned}$$

If $\theta_1 = \theta$ is taken as the reference angle and since $\cot\theta = \frac{m}{2D}$ from Equation 6.4, $\Delta\theta_1$ can be written as

$$\begin{aligned}\Delta\theta_1 &= \cot^{-1}\left[\frac{D}{d}\left(1+\frac{m^2}{4D^2}\right) + \frac{m}{2D}\right],\\ &= \cot^{-1}\left[\frac{1}{d}\left(\frac{4D+m^2}{4D}\right) + \frac{m}{2D}\right],\\ &= \cot^{-1}\left[\frac{1}{4D}\left((4D+m^2)\frac{1}{d} + 2m\right)\right]. \qquad (6.19)\end{aligned}$$

Equation 6.19 describes the change of angle $\Delta\theta_1$ as the sound source is passing and is dependent on $\Delta d = v\Delta t$ where $t = 0$, $\tau = 0$ at the reference point.

6.1.4 Time delay and vehicle velocity

The rate of change of τ as a function of t is very similar to the rate of change of θ expressed in Equation 6.19. An expression of τ is preferable to θ since τ is directly measurable from the microphone array TDOA techniques whereas θ must be first calculated. Therefore Equation 6.19 is used in this section to obtain τ as a function of t. At the reference point, τ equals 0. Before and after this moment the inter-microphone time delay τ changes as the sound source approaches, passes and grows distant. Assuming the vehicle is traveling at a constant velocity v, knowledge of the rate of change of τ and the geometrical equations can be used to estimate v. The following has been established:

1. Equation 6.11 has been obtained that describes the delay τ in terms of angle θ:
$$\tau = \frac{D}{c}\left[\frac{1}{\sin\theta_1} - \sqrt{1 + \left(\frac{m}{D} - \cot\theta_1\right)^2}\right]. \tag{6.20}$$

2. When the delay τ is 0, the angle θ can be described as $\cot\theta = \frac{m}{2D}$. This is taken as the reference point, where $t = \tau = 0$.

3. When the car moves from θ to a new angle $\theta_1 = \theta - \Delta\theta_1$, the change in angle is given in Equation 6.19 as
$$\Delta\theta_1 = \cot^{-1}\left[\frac{1}{4D}\left((4D + m^2)\frac{1}{d} + 2m\right)\right]. \tag{6.21}$$

Since $d = vt$, the delay τ in terms of time t can be found from 6.11 by setting $\theta_1 = \theta - \Delta\theta$ and taking Equation 6.19 into account. Starting with Equation 6.11:
$$\tau = \frac{D}{c}\left[\frac{1}{\sin\theta_1} - \sqrt{1 + \left(\frac{m}{D} - \cot\theta_1\right)^2}\right]. \tag{6.22}$$

Since $\cot\theta$ and $\cot(\Delta\theta_1)$ are known, it is useful to express everything in terms of $\cot\theta_1$. Therefore replace $1/\sin\theta_1$ in Equation 6.11 as follows:

$$\frac{1}{\sin\theta_1} = \pm\sqrt{1 + \cot^2\theta_1} \Rightarrow +\sqrt{1 + \cot^2\theta_1} \text{ since } 0 < \theta_1 < 180^o, \tag{6.23}$$
$$\tau = \frac{D}{c}\left[\sqrt{1 + \cot^2\theta_1} - \sqrt{1 + \left(\frac{m}{D} - \cot\theta_1\right)^2}\right]. \tag{6.24}$$

In Equation 6.24 $\cot\theta_1$ can be replaced once a suitable expression is derived as follows:

$$\begin{aligned} \cot\theta_1 &= \cot(\theta - \Delta\theta), \\ &= \frac{\cot\theta\cot\Delta\theta + 1}{\cot\Delta\theta - \cot\theta}. \end{aligned}$$

But since it is known that $\cot\theta = \frac{m}{2D}$, this becomes

$$\cot\theta_1 \quad = \quad \frac{\frac{m}{2D}\cot\Delta\theta + 1}{\cot\Delta\theta - \frac{m}{2D}}.$$

Also, it is known from Equation 6.19 that $\Delta\theta_1 = \cot^{-1}\left[\frac{1}{4D}\left((4D+m^2)\frac{1}{d} + 2m\right)\right]$. So the following can be written:

$$\begin{aligned}
\cot\theta_1 &= \frac{\frac{m}{2D}\frac{1}{4D}\left[(4D+m^2)\frac{1}{d} + 2m\right] + 1}{\frac{1}{4D}\left[(4D+m^2)\frac{1}{d} + 2m\right] - \frac{m}{2D}}, \\
&= \frac{[m(4D+m^2)\frac{1}{d} + 2m^2 + 8D^2]/8D^2}{[(4D+m^2)\frac{1}{d} + 2m - 2m]/4D}, \\
&= \frac{1}{2D}\frac{m(4D+m^2)\frac{1}{d} + 2m^2 + 8D^2}{(4D+m^2)\frac{1}{d}}, \\
&= \frac{1}{2D}\left[m + \frac{2(m^2+4D^2)}{(4D+m^2)}d\right], \\
&= \frac{m}{2D} + \frac{m^2+4D^2}{m^2+4D}\frac{d}{D}. \qquad (6.25)
\end{aligned}$$

Rearranging Equation 6.25 gives the following equation describing d in terms of angle θ_1:

$$d = D\left(\frac{m^2+4D}{m^2+4D^2}\right)\left(\cot\theta_1 - \frac{m}{2D}\right). \qquad (6.26)$$

Using the expression for $\cot\theta$ in Equation 6.25, it is possible to rewrite the description of Equation 6.24 and remove all references to angles θ, θ_1, θ_1' or $\Delta\theta_1$ as follows:

$$\begin{aligned}
\tau &= \frac{D}{c}\left[\sqrt{1 + \left(\frac{m}{2D} + \frac{m^2+4D^2}{m^2+4D}\frac{d}{D}\right)^2} - \sqrt{1 + \left(\frac{m}{D} - \frac{m}{2D} - \frac{m^2+4D^2}{m^2+4D}\frac{d}{D}\right)^2}\right], \\
&= \frac{D}{c}\left[\sqrt{1 + \left(\frac{m}{2D} + \frac{m^2+4D^2}{m^2+4D}\frac{d}{D}\right)^2} - \sqrt{1 + \left(\frac{m}{2D} - \frac{m^2+4D^2}{m^2+4D}\frac{d}{D}\right)^2}\right]. \qquad (6.27)
\end{aligned}$$

Equation 6.27 describes the inter-microphone time delay τ as a function of vehicle distance traveled based on the known geometry and reference point where $\tau = 0$ with constant velocity. Further simplifications are possible, depending on the relationships between D, M, i.e. $m \ll D \Rightarrow \frac{m}{D} \ll 1$.

6.1.5 Road length and time delay

An expression of d as a function of τ is now sought. This can be readily obtained by re-arranging Equation 6.27, repeated below.

$$\tau \;=\; \frac{D}{c}\left[\sqrt{1+\left(\frac{m}{2D}+\frac{m^2+4D^2}{m^2+4D}\frac{d}{D}\right)^2}-\sqrt{1+\left(\frac{m}{2D}-\frac{m^2+4D^2}{m^2+4D}\frac{d}{D}\right)^2}\right] \tag{6.28}$$

Substitution is used to simplify Equation 6.27, where $a = \frac{m^2+4D^2}{m^2+4D}\frac{1}{D}$, $b = \frac{D}{c}$ and $e = \frac{m}{2D}$.

$$\tau = b\left[\sqrt{1+(e+ad)^2}-\sqrt{1+(e-ad)^2}\right], \tag{6.29}$$

$$\frac{\tau}{b} = \sqrt{1+(e+ad)^2}-\sqrt{1+(e-ad)^2}, \tag{6.30}$$

$$\frac{\tau}{b}+\sqrt{1+(e-ad)^2} = \sqrt{1+(e+ad)^2}, \tag{6.31}$$

$$\frac{\tau^2}{b^2}+1+(e-ad)^2-2\sqrt{\frac{\tau^2}{b^2}}\sqrt{1+(e-ad)^2} = 1+(e+ad)^2, \tag{6.32}$$

$$\frac{\tau^2}{b^2}+1+(e-ad)^2-\frac{2\tau}{b}\sqrt{1+(e-ad)^2}-1-(e+ad)^2 = 0, \tag{6.33}$$

$$\frac{\frac{\tau^2}{b^2}+(e-ad)^2-(e+ad)^2}{\frac{2\tau}{b}} = \sqrt{1+(e-ad)^2}, \tag{6.34}$$

$$\frac{\frac{\tau^2}{b^2}+e^2+a^2d^2-2aed-[e^2+a^2d^2+2aed]}{\frac{2\tau}{b}} = \sqrt{1+(e-ad)^2}, \tag{6.35}$$

$$\frac{b\left[\frac{\tau^2}{b^2}-4aed\right]}{2\tau} = \sqrt{1+(e-ad)^2}, \tag{6.36}$$

$$\frac{\frac{\tau^2}{b}-4abed}{2\tau} = \sqrt{1+(e-ad)^2}, \tag{6.37}$$

$$\left[\frac{\tau}{2b}-\frac{2abed}{\tau}\right]^2 = \left[\sqrt{1+(e-ad)^2}\right]^2, \tag{6.38}$$

$$\frac{\tau^2}{4b^2}+\frac{4(abed)^2}{\tau^2}-\frac{4abed\tau}{2b\tau} = 1+(e-ad)^2, \tag{6.39}$$

$$\frac{\tau^2}{4b^2}+\frac{4(abed)^2}{\tau^2}-2aed = 1+e^2+(ad)^2-2aed, \tag{6.40}$$

$$\frac{\tau^2}{4b^2} + \frac{4(abcd)^2}{\tau^2} - 1 - e^2 - (ad)^2 = 0, \tag{6.41}$$

$$d^2\left[\frac{4(abe)^2}{\tau^2} - a^2\right] = 1 + e^2 - \frac{\tau^2}{4b^2}, \tag{6.42}$$

$$d^2 = \frac{1 + e^2 - \frac{\tau^2}{4b^2}}{\frac{4(abe)^2}{\tau^2} - a^2}. \tag{6.43}$$

Now that d is written as a function of everything else, lets replace the substitutions from earlier, where $a = \frac{m^2+4D^2}{m^2+4D}\frac{1}{D}$, $b = \frac{D}{c}$ and $e = \frac{m}{2D}$:

$$d = \sqrt{\frac{1 + \frac{m^2}{4D^2} - \frac{\tau^2}{\frac{4D^2}{c^2}}}{\left(\frac{1}{\tau}\frac{4D^2}{c^2}\frac{m^2}{4D^2}\right)\left(\frac{m^2+4D^2}{m^2+4D}\frac{1}{D}\right)^2 - \frac{(m^2+4D^2)^2}{(m^2+4D)^2}\frac{1}{D^2}}}, \tag{6.44}$$

$$d = \sqrt{\frac{1 + \frac{m^2}{4D^2} - \frac{\tau^2 c^2}{4D^2}}{\frac{m^2}{\tau c^2}\left(\frac{m^2+4D^2}{m^2+4D}\frac{1}{D}\right)^2 - \left(\frac{m^2+4D^2}{m^2+4D}\frac{1}{D}\right)^2}}. \tag{6.45}$$

6.1.6 Summary of relevant equations

Table 6.1: Generalized equations modelling a moving sound source

Equation 6.5	$m = D(\cot\theta_1 + \cot\theta_2)$
Equation 6.24	$\tau = \frac{D}{c}\left[\sqrt{1+\cot^2\theta_1} - \sqrt{1+\left(\frac{m}{D} - \cot\theta_1\right)^2}\right]$
Equation 6.26	$d = D\left(\frac{m^2+4D}{m^2+4D^2}\right)\left(\cot\theta_1 - \frac{m}{2D}\right)$
Equation 6.27	$\tau = \frac{D}{c}\left[\sqrt{1+\left(\frac{m}{2D} + \frac{m^2+4D^2}{m^2+4D}\frac{d}{D}\right)^2} - \sqrt{1+\left(\frac{m}{2D} - \frac{m^2+4D^2}{m^2+4D}\frac{d}{D}\right)^2}\right]$
Equation 6.45	$d = \sqrt{\frac{1+\frac{m^2}{4D^2}-\frac{\tau^2 c^2}{4D^2}}{\frac{m^2}{\tau c^2}\left(\frac{m^2+4D^2}{m^2+4D}\frac{1}{D}\right)^2-\left(\frac{m^2+4D^2}{m^2+4D}\frac{1}{D}\right)^2}}$

A summary of the relevant equations is shown in Table 6.1. These equations can be used to simulate and model the behaviour and change in parameters due to a moving source. This is described in Section 6.2.

6.2 Description of system parameters

Parameter value choice determines the performance of the described audio traffic monitoring system. This section describes the relationship between parameter values to achieve a range of desired but conflicting targets of system speed and accuracy. System performance is determined for a variety of parameter values by using the mathematical model derived in the previous section to simulate a moving sound source.

Table 6.2: Audio traffic monitoring system parameters

Geometrical parameters	τ	inter-microphone time delay
	θ	observation angle
	d	source road distance from reference point
	m	inter-microphone distance
	D	distance to the road centre
	v	sound source road velocity
Signal Processing parameters	f_s	sampling frequency
	L_w	window length
		window shape
	O_w	hop size

The system parameters are listed in Table 6.2 and illustrated in Figure 6.5. A description of the system is as follows. Two microphones are placed a distance m apart, parallel to the road. The distance from the microphone pair to the centre of the road is noted as D. A sound source or vehicle is assumed to pass the microphone array while travelling along the road at velocity v. The source is observed by the microphone pair at an angle θ when in range. This angle may be determined by calculating the inter-microphone time delay τ. The microphone signals are cross-correlated to calculate τ.

Since the microphone signals are discretised at sampling frequency f_s, there is a

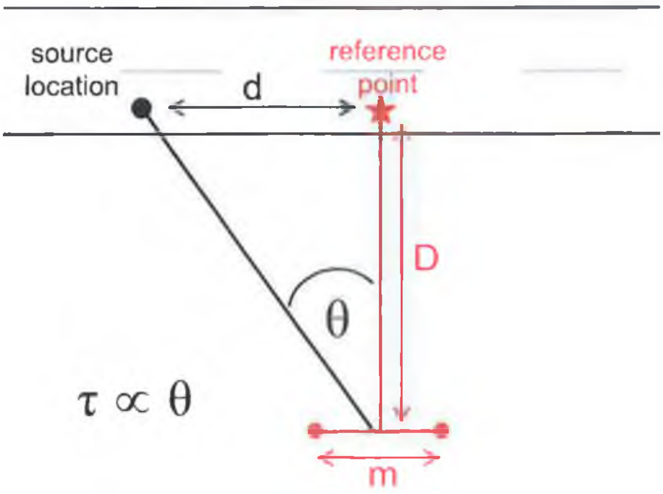


Figure 6.5: Simple road geometry depicting system parameters

corresponding discrete series of measurable values for τ ranging from zero to τ_{max}, with the smallest measurable interval being τ_{min}. As the vehicle moves, its location along the road can be described as a distance d from the reference point opposite the microphone pair where $\tau = 0$. Sections or windows of the microphones signals are cross-correlated to obtain τ, where the *window length* and *shape* are relevant parameters. To retain some degree of smoothness, successive analysis windows overlap previous ones slightly, where the jump or *hop size* dictates the number of samples to progress for each iteration. For the remainder of this section each system parameter is described individually.

6.2.1 Distance between microphones

The inter-microphone distance parameter m is very important as it influences system accuracy and is a key parameter in dictating the shape of the moving source model. The further apart the microphones are, the greater the maximum measurable time delay τ_{max} will be, making it easier to distinguish different location. Recall that τ is the time a sound requires to propagate the extra distance to the further microphone. This extra distance can only be less than or equal to the distance m between the microphones. The time taken to traverse such a distance depends on the speed of propagation, in this case the speed of sound c. The largest measurable time delay between two adjacent microphones a distance m apart is τ_{max} and occurs when $\theta = 0$.

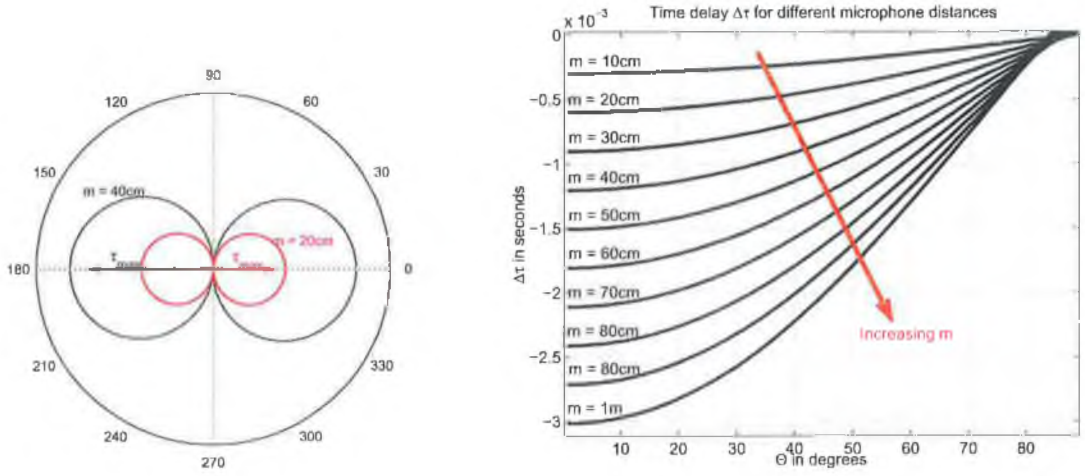


Figure 6.6: (a) Cyclical change in magnitude of time delay τ from 0 to τ_{max} as a function of observation angle, illustrated for inter-microphone values m=20 and 40cm (b) τ versus θ_1 for different m

Since $\cot\theta = 0$, Equation 6.24 can be used to describe τ_{max} as follows:

$$\begin{aligned}\tau &= \frac{D}{c}\left[\sqrt{1+\cot^2\theta_1} - \sqrt{1+\left(\frac{m}{D}-\cot\theta_1\right)^2}\right],\\ \Rightarrow \tau_{max} &= \frac{m}{c}.\end{aligned} \qquad (6.46)$$

Figure 6.6(a) displays the cyclical variation in the magnitude of τ as the observation angle is rotated around 360^o degrees, where $0 \leq \tau \leq \tau_{max}$. In the first quadrant (0-90^o), τ reduces from τ_{max} to 0 as the observation angle increases. From 90^o to 180^o, τ increases in magnitude from 0 to -τ_{max}. This is symmetrically replicated between 180 and 360^o. Figure 6.6(a) also illustrates how τ_{max} is doubled in value by a doubling of the value of m. The ranges of τ values are plotted for when m = 20cm and 40cm respectively. The influence of the value of m is further illustrated in Figure 6.6(b), where the relationship between τ and θ changes for different values of m.

The maximum value of m is constrained by the requirement from Section 3.2.1 that $D \gg m$ in order to to enable a far-field assumption of the received signals. Furthermore, as the microphones move further apart, there is a greater probability that the audio signals they receive increase in independence and become less correlated. This makes the cross-correlation based source localization task increasingly difficult to the point where it is impossible. Since the system is based on measuring cross-correlation to determine source location, the microphones are placed in close proximity.

6.2.2 Distance to the road

D is the distance between the centre of the road and the microphone array as illustrated in Figure 6.5. The value of D determines the level of source sound attenuation of the received signal (as described in Section 3.2.1), the maximum observable road surface and the distinguishable road locations for a single m value. An appropriate value for D should satisfy the following criteria:

1. The sound received at the microphone array is not attenuated to an excessive level;

2. To enable a far-field assumption, D should be $\gg m$. As described in Section 3.2.1, once a far-field scenario can be assumed the received signals may be considered as plane waves with a single propagating direction;

3. The observable road surface is sufficiently long to obtain a measurable evaluation of vehicle behaviour.

Changes in the value of D have a negligible effect on the moving source model $\tau(t)$, in other words the model is unaffected by the choice of value for D. Therefore the only constraints in selecting D are that the sound attenuation is not excessive, and $D \gg m$. D was set at values between 1 and 5m, depending on the recording location.

6.2.3 Sampling frequency

The sampling frequency f_s is a measure of how often a continuous signal is sampled. The sampling frequency determines the precision of τ, since

$$\Delta\tau_{min} = \frac{1}{f_s}. \tag{6.47}$$

If the sound source is moving at velocity v, there is a limited time when it is within range of the array. Projecting θ onto the road surface results in a series of discrete possible source locations. The location resolution is dictated by the sampling frequency f_s but further limited by vehicle velocity v, since a fast-moving vehicle will change location between the sampling times. Therefore, the measured source distance d will have a location error that depends on θ, f_s and D. This is even true for an ideal situation that discounts the retardation effect introduced in Section 4.1.2.

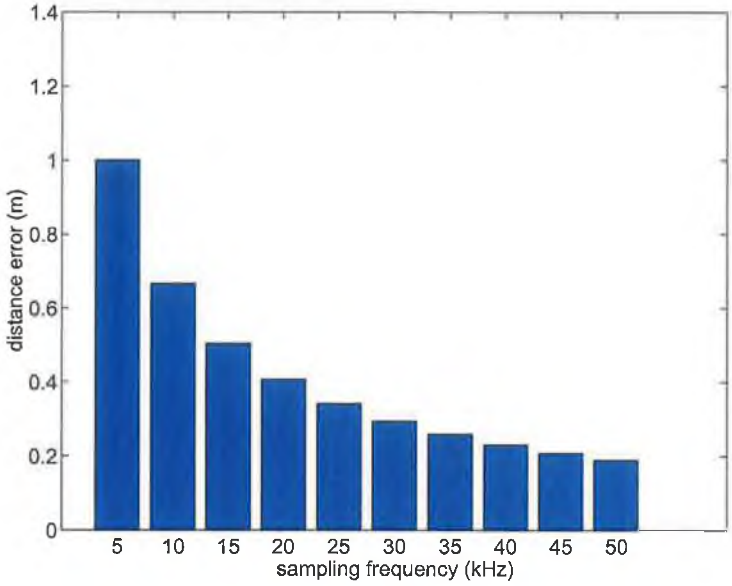


Figure 6.7: f_s versus quantized error of distance d

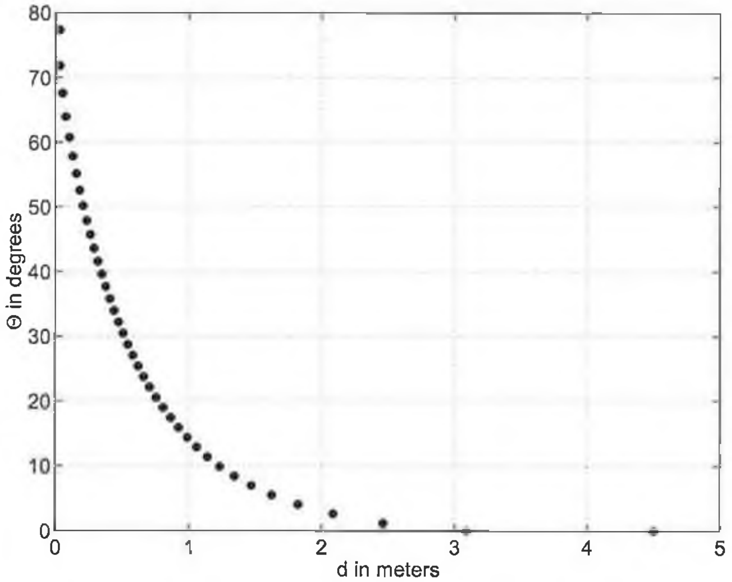


Figure 6.8: Discrete d versus θ_1 at a sampling frequency of 44.1kHz

Figure 6.7 graphs the error in distance d relative to a series of sampling frequencies. Equation 6.26 can be used to illustrate the effect of changes in θ on d, shown in Figure 6.8. From both diagrams, it can be seen that the choice of sampling frequency dictates the source location accuracy.

The sampling frequency also determines the signal bandwidth, where the highest measurable frequency is half the sampling frequency for a signal, called the Nyquist frequency [137]. In order to maximize the signal data available for experiments, a sampling frequency of 44.1kHz was used for recording traffic data. Theoretically the measurable frequency bandwidth is 0 to 22.05kHz, assuming the microphones and hardware are capable of accurately measuring this bandwidth. The choice of sampling frequency has an influence on processing speed. This relationship is described in Section 8.5.1 based on measurements applying different analysis techniques to traffic data.

Interpolation of the audio or cross-correlation data can be performed to artificially increase the sampling frequency, as described in Section 5.3.4. Interpolation results in an increase of the time-delay resolution, reducing the required sampling frequency for a particular accuracy. Interpolating by a factor of 4 raises the effective sampling frequency to 176.4kHz. The choice of sampling frequency has an influence on processing speed. This relationship is described in Section 8.5.1 based on measurements applying different analysis techniques to traffic data.

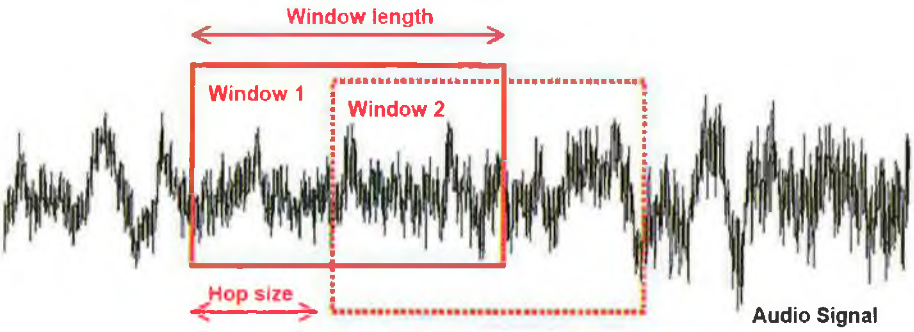


Figure 6.9: Window length and hop size

6.2.4 Window size

A window is applied to isolate a section of the microphone signals prior to signal analysis. The chosen window can be described by its length L_w, shape and successive overlap O_w as illustrated in Figure 6.9. The choice of window length has an impact on the performance of signal processing approaches. Constraints on L_w are summarized below:

- to satisfy signal processing assumptions, the window size must be small enough for the signals to be considered stationary;
- The window must be long enough to obtain a reliable result;
- The sound source should not have moved significantly within a window, in order to be able to pinpoint the location with acceptable accuracy.

One purpose of the window is to ensure the spectral characteristics are reasonably stationary over the duration of the window, since stationarity is a requirement for the cross-correlation method implemented. The more rapidly the signal characteristics change, the shorter the window should be. As the L_w becomes smaller, frequency resolution decreases. On the other hand, as L_w decreases, the ability to resolve temporal changes increases. Consequently, the choice of L_w becomes a trade-off between frequency resolution and time resolution with stationarity an added issue. This is sometimes called the spectral-temporal resolution trade-off. As described in Section 5.1.1, acoustic signals are rarely absolutely stationary. Fourier transform and cross-correlation assume the signals are stationary or at least wide-sense stationary, where *stationary* signals are constant in their statistical parameters over time [20].

Knowledge of the expected signal properties are required to decide in what time duration the received signal can be considered stationary. Stationarity tests were performed on the audio signal to determine an appropriate value for L_W. Unfortunately, an appropriate window size that achieved wide-sense signal stationarity could not be defined for the recorded audio traffic signals, since all window sizes resulted in an excessively large variation in statistical characteristics. This was true for any size of window. However, both the cross-correlation sequence and Fourier transform methods performed as expected, despite the stationarity assumption not being satisfied. Therefore a 0.11s window was chosen for experimental purposes. This value was found to provide good results.

6.2.5 Window shape

In addition to choosing the length of the window, an appropriate windowing function or shape should be determined. One approach is to simply use a rectangular window. A signal constructed in this manner has sharp discontinuities at its edges. The frequency-domain version of the signal consists of a main lobe and series of large side lobes which result in some undesirable ringing effects in the frequency response. These undesirable effects are best alleviated by the use of windows that do not contain abrupt discontinuities in their time-domain characteristics and have correspondingly low sidelobes in their frequency-domain characteristics. Some of the commonly used window sequences are shown in Figure 6.10, which are symmetric about the time (N-1)/2 [50, 156]. Windows such as the Kaiser, Hamming, Hanning and Blackman tend to distort the temporal waveform over the range of N points, but with the benefit of less abrupt truncations at the boundaries. The popular Hamming window was chosen, which attenuates the side-lobes by 30dB. Equation 6.48 represents a Hamming window.

$$w(N) = 0.54 - 0.46 \cos \frac{2\pi n}{N-1}. \tag{6.48}$$

6.2.6 Window overlap

Overlapping windows are often used for a smoother transition from window to window. Sometimes called the hop size, the window overlap O_W is used to describe the amount by which the analysis time origin is advanced for each successive window, as illustrated in Figure 6.9. A smaller overlap will give more analysis points and therefore smoother results across time, but the computational expense is proportionately greater. The minimum overlap has a lower constraint due to the sampling frequency. Since the audio signals are discrete, the smallest possible overlap is a single sample or a time duration of $\frac{1}{fs}$ seconds. Using such a value is unrealistic, as there is undue repetition and no gain in accuracy.

The overlap should be small enough so any *measurable* change in source location is always captured. The source location can only be measured at sampled time delay intervals τ with an quantization factor determined by the sampling frequency and interpolation level. To determine a suitable overlap, the smallest measurable changes in τ must first be obtained. The faster the source velocity, the less time a source takes to pass the microphone array. The largest change in τ occurs when the source

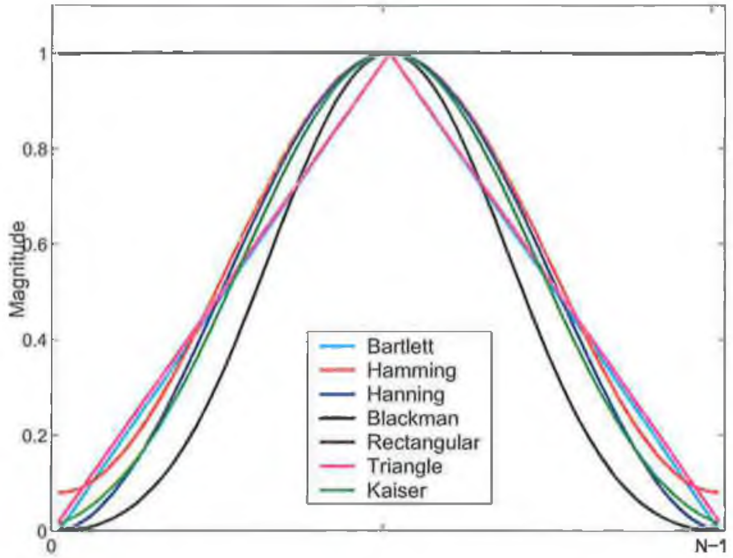


Figure 6.10: Shapes of several window functions

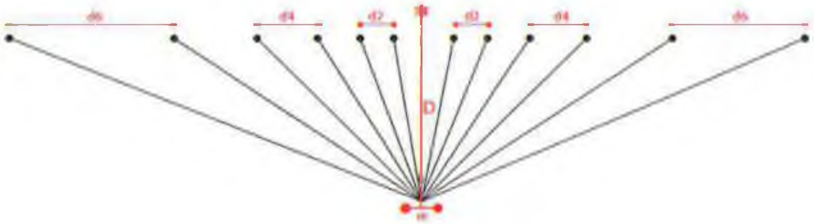


Figure 6.11: Road geometry discrete τ

is opposite the microphone array, i.e. around when $\tau = 0$. The smallest necessary overlap can therefore be determined by calculating the smallest amount of time a source moving at the highest possible velocity takes to traverse a measurable discrete $\Delta\tau$ for the full range of possible microphone array geometries.

Figure 6.11 provides an illustration of the discrete measurable $\Delta\tau$ values, represented by the series of radial lines. It is possible to calculate the values of τ for a range of source velocities using the model equations from Chapter 6, thereby determining the required time resolution. The largest possible overlap clearly cannot exceed the window size. This presents a range of possible values to choose from. For a sampling frequency of 44.1kHz, the smallest measurable $\Delta\tau$ around the reference point is $22.67\mu s$. The microphone array geometry values m and D have an influence on the measurable distances but not on what value of $\Delta\tau$ can be measured. Calculations were made using a range of values, where $1 < v \leqslant 300$km/h, $0.05 < m \leqslant 5$m, $0.5 < D \leqslant 20$m, $2 < f_s \leqslant 44.1$kHz. For a sampling frequency of 44.1kHz when $v = 300$km/h, it takes 14 of the smallest possible time increments to traverse the smallest measurable τ.Therefore, for the same sampling frequency, the overlap need not be less than $14 \times \frac{1}{fs}$ or 3.17×10^{-4} seconds in duration.

6.2.7 Observation angle

The location of a source can be described using polar coordinates, which consist of two variables; the source angle relative to a reference point θ and the distance to the source. In a far-field scenario (as introduced in Section 3.2.1) it is not possible to determine the distance to the source with a microphone pair, therefore only angle θ is considered. However it is not θ that is measured using a TDOA localization approach, but rather the inter-microphone time delay τ. The relationship between inter-microphone time delay τ and source observation angle θ_1 is described in Equation 6.24 and visualised in Figure 6.12. It can be seen from the graph that for small time delays the angle is linear, but for very large time delays a nonlinearity between τ and θ becomes stronger when approaching the limit τ_{max}. For the case of Figure 6.12, m is set at 20cm, D is 10m and c is estimated as 331.1m/s over 90 degrees.

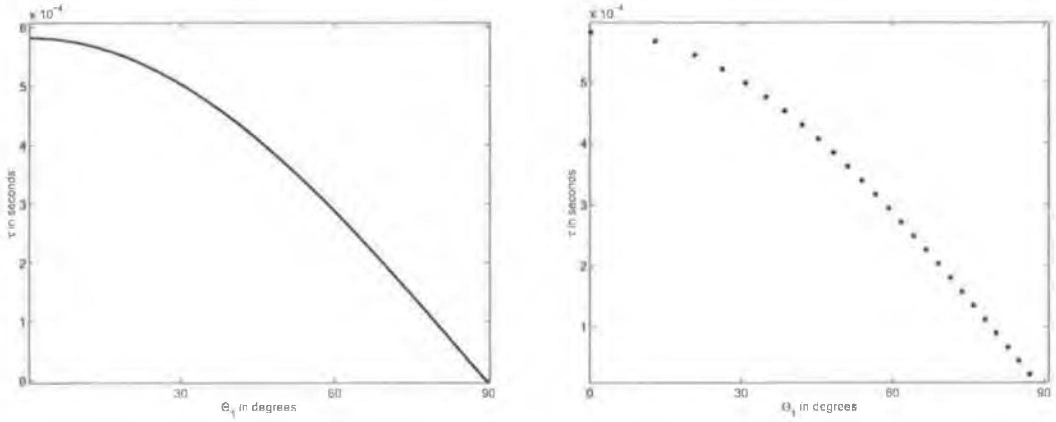


Figure 6.12: τ versus θ_1 for a (a) continuous signal (b) discrete signal at a sampling frequency of 441.kHz

6.3 Theoretical system performance

The performance of the audio-based traffic monitoring system can be simulated using the mathematical model derived in Section 6 and an understanding of the parameters described in Section 6.2. Instead of specifying suitable parameter values and then calculating the accuracy, this section specifies system accuracy and calculates the necessary parameter values to achieve such an accuracy. In this manner, the number and accuracy of vehicle measurements can be predicted for a given set of parameter values. The number of possible measurements, observable road length and theoretical velocity accuracy are described.

6.3.1 Number of possible measurements

The number and range of time-delay measurements available to the microphone array is limited by system parameters. In order to evaluate the maximum number of time-delay measurements, a range of parameter values are examined. As described in Section 6.2, τ ranges from 0 to τ_{max} in increments of size τ_{min}. Therefore the maximum number of possible τ measurements, N_τ, can be described as

$$N_\tau \leq \frac{\tau_{max}}{\tau_{min}} = \frac{mf_s}{c}. \tag{6.49}$$

The parameters f_s and m determine the value of τ_{min} and τ_{max} respectively, thereby influencing the maximum number of τ measurements, N_τ. Figure 6.13 illustrates N_τ constrained by f_s and m. N_τ is not the number of measurements obtained for all sources, simply the maximum number of possible τ measurements. As expected, it can be observed that a larger number of measurements are possible for a larger distance between microphones. Similarly, the number of measurements increases for higher sampling frequencies. Interpolation is not taken into account for this graph. For a inter-microphone distance of 20cm with a sampling frequency of 44.1kHz, the number of possible measurements for τ is approximately 30 for angles ranging between -45 and 45 degrees.

Figure 6.14 illustrates the theoretical impact of changing the distance to the road, D. The sampling frequency is 44.1kHz and the range of observation angles are between -45 and 45 degrees. The number of measurable values of τ are plotted for a range of values of m and values of D. As may be observed, the value of D has a negligible influence on the number of available measurements, provided $m \ll D$. Therefore, D can be disregarded when choosing parameters to obtain a particular number of measurements. It is only necessary to consider D once m is chosen, provided D is not so large that the sound is excessively attenuated.

The velocity of the vehicle has not been considered up to this point when determining the number of τ measurements. A moving source may traverse the observed road length at a high velocity such that it is undetected in some of the location bins, unless the sampling frequency is sufficiently high. Figure 6.15 illustrates the maximum number of measurements for a moving source at different velocities and for a range of sampling frequencies. The number is based on observing a road surface between the angles of -45 and 45 degrees, where the inter-microphone distance m is 20cm. It is assumed that the time shift of successive windows is a single sample. It can be observed from the graph that an extremely high number of measurements are available for sources with low velocities, particularly below 50 km/h. As the source velocity increases, the maximum number of measurements initially decreases rapidly then at a slower rate. The maximum number of τ measurements have been displayed in this graph, as opposed to the range of measured τ values. The range and accuracy of measurements for a moving source at different velocities are described in Section 6.3.3.

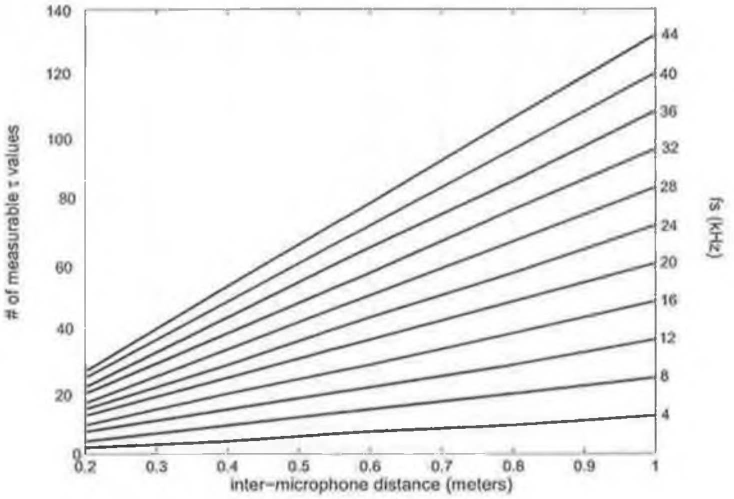


Figure 6.13: Maximum number of observations (measurable time delay values) between -45 and 45 degrees, constrained by the sampling frequency and inter-microphone distance

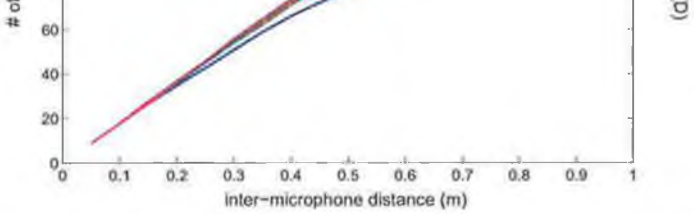


Figure 6.14: Maximum number of observations between -45 and 45 degrees for a selection of inter-microphone distances (m) and distances to the road (D), where fs = 44.1kHz, c = 331.1m/s

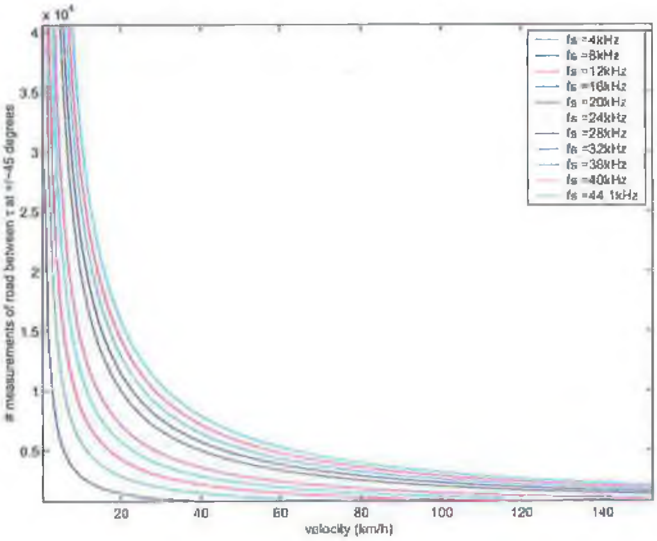


Figure 6.15: Maximum number of observations between -45 and 45 degrees for increasing velocity and a range of sampling frequencies, where m=0.2m

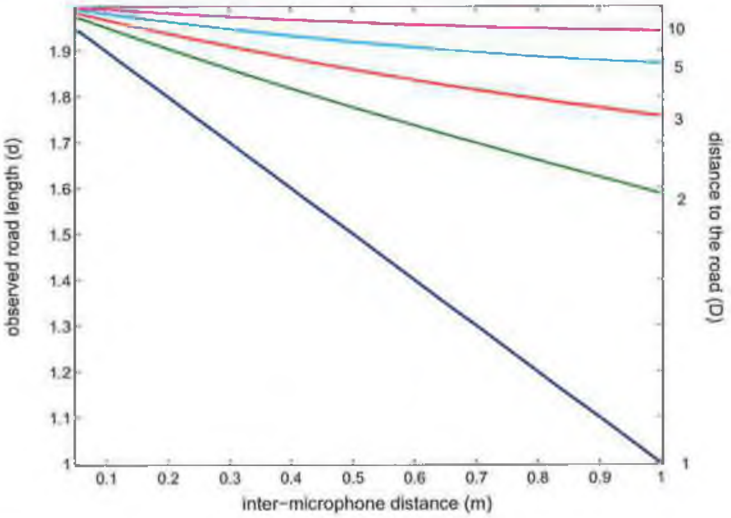


Figure 6.16: Observed road length between -45 and 45 degrees for a selection of inter-microphone distances (m) and distances to the road (D), where fs = 44.1kHz, c = 331.1m/s,

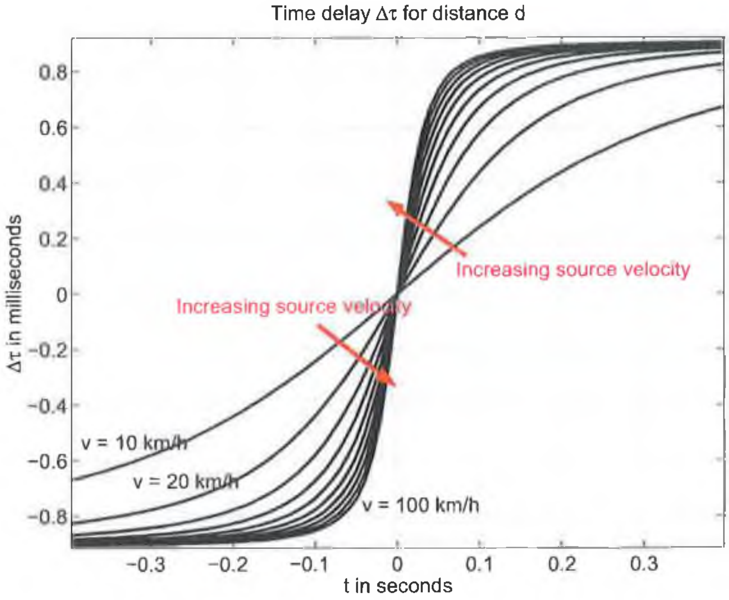


Figure 6.17: τ versus t for different v

6.3.2 Road length

The length of road over which measurements are obtained was not considered in Section 6.3.1. It is important to know this road length, as it places constraints on the number of sound sources that could be simultaneously present within the observed road length. The road length observed by the microphone array depend on the values of m and D.

Figure 6.16 graphs the observed road length for a selection of inter-microphone distances relative to the distance to the road. The observed road length is visibly larger when the distance to the road is greater. Furthermore, as the distance between microphones increases, the observed road length decreases linearly. When m is 20cm and the distance to the road is 2m, the observed road length between -45 and 45 degrees is just over 1.9m. From Figure 6.13, and based on a sampling frequency of 44.1kHz, the number of possible measurements over the road distance of 1.9m is approximately 30 for the defined parameter values. In the case of Figure 6.17(a), a road length of approximately 6-10m is observable with reasonable accuracy, when $c = 331.1$, $D = 7$m and $m = 0.3$m. The measurement graphs may be consulted in this manner to cross-check the impact of a particular parameter value.

6.3.3 Theoretical accuracy of velocity estimation

Consider Figure 6.17, in which a number of moving source models with different velocities are presented. It can be observed that the differences between successive models becomes smaller for higher velocity values. However, if the differences between models are smaller than the measured time resolution, the accuracy of matching the model and thereby determining velocity is compromised. The method used to estimate vehicle velocity is to determine the velocity parameter of the best-fitting model. Particularly for data from vehicles travelling at high velocities, the difference between models is significantly reduced. The time resolution of the data, and therefore accuracy limits are determined by the audio signal sampling frequency. The following section quantifies the time resolution required to distinguish velocities to a range of accuracies.

Time resolution required to distinguish velocities to a range of accuracies

The theoretical accuracy of vehicle velocity estimation is now quantified as well as the precision of measured time required to achieve such an accuracy. The velocity of a moving source is determined from the parameters of the best-matching moving source model $\tau(t, v)$. Since the model $\tau(t, v)$ is a series of discrete time measurements, the difference between two such models is time-based. If the distinction between two models is finer than the sampling rate f_s, it is impossible to differentiate between the two models. Therefore the accuracy in specifying the velocity of a model is based on the measurable time difference between models. In order to quantify the time precision requirements to achieve specific velocity accuracies, an array is generated that describes the largest difference between a reference model and test model models for a range of velocities.

Consider a reference moving source model $\tau_{ref}(t, v)$ with a constant velocity v_i. Consider also a test model $\tau_{test}(t, v)$ with a different constant velocity, where the difference between the model velocities is Δv. The two models are compared and the largest difference between them stored in matrix A, described in Equation 6.50.

$$A(v, \Delta v) = max\left[\tau_{ref}(t, v_i) - \tau_{test}(t, \Delta v_j)\right], \qquad \text{for } v_i = v_{min} : v_{max} \quad \text{and } \Delta v_j = 1 : v_{max}. \tag{6.50}$$

Multiple test models are compared against the reference model $\tau_{ref}(t, v_i)$ for a range of test velocities from $\Delta v = v_{min}$ to v_{max}. This is in order to establish time difference values for increasing Δv. Since the reference model is non-linear and velocity-dependent, the time difference between models depends not only on Δv but also the actual value of v_i, i.e. for higher values of v_i and equivalent Δv, the maximum time difference is lower. Therefore A represents the maximum inter-model time difference for a range of velocity values from v_{min} to v_{max} in one dimension, and velocity accuracies Δv from 1 to v_{max} in the other dimension.

If the minimum measurable time step τ_{min} is smaller than all values of A, there is no difficulty in distinguishing all velocities to the maximum displayed accuracy. However, if the minimum measurable time step τ_{min} exceeds some or all values of A, then not every velocity accuracy represented by A is achievable with the particular sampling frequency that determined τ_{min}. Either the tolerable velocity accuracy or the specified sampling frequency must be compromised.

Figure 6.18 illustrates the time precision required to distinguish vehicle velocities on the y-axis, while the x-axis represents vehicle velocity ($vRef$). Each curve presents a particular velocity accuracy (Δv) from 1km/h to 50km/h. The time resolution required for a particular velocity accuracy decreases as the velocity increases. Overlayed on the graph are horizontal lines representing the minimum time resolution available due to certain sampling frequencies (4, 8, 12, 20 and 44kHz respectively). This diagram confirms two important points:

- higher velocities require a higher time precision to measure velocity to the same accuracy;
- typical audio sampling frequencies (2-44.1kHz) do not achieve the time precision required to estimate velocity of vehicles travelling at a speed to be expected (0-250km/h) to a tolerable accuracy (±10km/h).

Figure 6.19 further quantifies the relationship between time precision and accuracy in velocity measurement. In this diagram, the velocity accuracy achieved with a particular time precision is displayed, where it is assumed that a vehicle is travelling between 1 and 250km/h.

An interpolation rate of 4 was applied to the cross-correlation sequence to artificially increase the sampling frequency, as described in Section 5.3.4. This reduces the required sampling frequency for a particular accuracy by a factor of 4. The

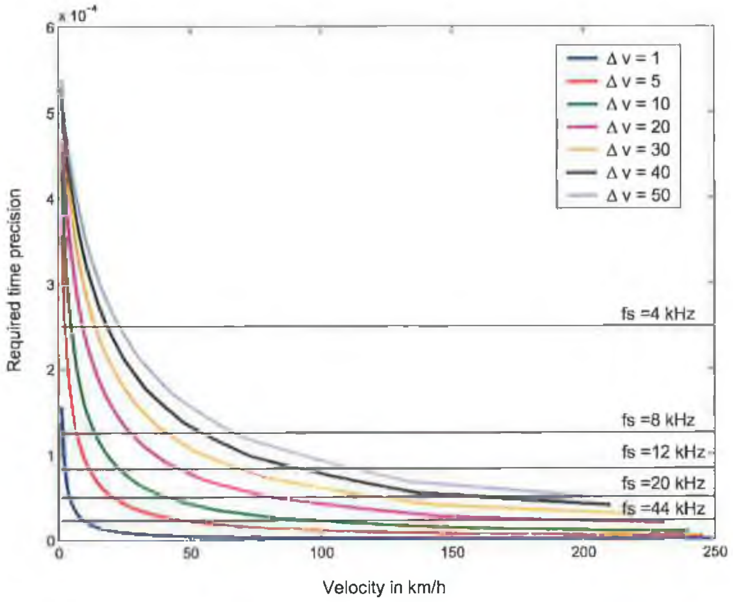


Figure 6.18: Required time precision to distinguish increasing vehicle velocities based on the moving source model. Each Δv represents a particular velocity accuracy from 1km/h to 50km/h. Each horizontal black line denotes the time precision resulting from that particular sampling frequency, illustrated for 4, 8, 12, 20 and 44kHz

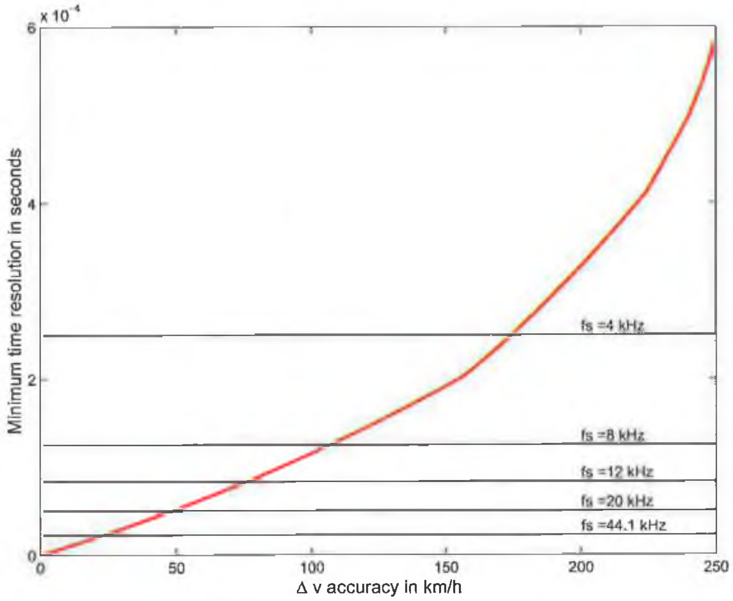


Figure 6.19: Minimum time resolution required to attain velocity accuracy Δv for a vehicle travelling within the range 1 to 250km/h. Each horizontal black line denotes the time precision resulting from that particular sampling frequency, illustrated for 4, 8, 12, 20 and 44kHz

interpolation raises the effective sampling frequency to 176.4kHz, meaning that the precision in measuring velocity is within an accuracy of ± 5.88 km/h.

This section has described a calculation of the required sampling frequency to achieve particular velocity accuracies. These accuracies are purely theoretical and do not take noisy data, cross-correlation errors or pattern analysis errors into account. Nevertheless, it is clear that very small time measurements are required in order to measure velocity to a high accuracy. This places a constraint on the minimum sampling frequency. Section 8.4.2 describes experimental results to measure vehicle velocity based on real traffic data with a sampling frequency of 44.1kHz.

Acceleration of moving source

The model describing a source location assumes the moving source has a constant velocity. In reality, vehicles may be accelerating or decelerating, resulting in velocity uncertainty. This uncertainty can be calculated based on knowledge of the maximum acceleration/deceleration of a vehicle combined with knowledge of the road distance under observation.

6.4 Summary of system parameters and accuracy

During this chapter, equations have been derived to describe the geometrical relationships between a microphone array adjacent to the road and a moving sound source. These equations were used to simulate a moving source based on a range of parameter values. In this manner, system accuracy and performance can be evaluated for a given set of parameter values. In some cases a trade-off must be made to balance conflicting priorities. Depending on the resources available and system requirements, the chosen parameter values may vary. This is the primary reason for describing the implications of a range of parameter values.

For the purposes of experiments on real traffic data described in Chapter 8, parameter values were restricted to a single or two different values. The parameter values used by the automatic traffic monitoring system are listed in Table 6.3.

When m is 0.2m and $D < 10$m, the observed road length is a maximum of 6m, or 2m between -45 and 45 degrees, as illustrated in Figure 6.16. The average length of

Table 6.3: Experimental audio traffic monitoring system parameters

m	inter-microphone distance	0.1 to 0.2m
D	distance to the road centre	0.5 to 5m
f_s	sampling frequency	44.1kHz
	interpolation factor	4
L_w	window length	0.1s
	window shape	Hamming
O_w	window overlap	22ms

a typical car is 4.25m, therefore it is unlikely that more than a single vehicle could occupy the observed road length in a single lane at the same time. Therefore, for a road with two lanes, the maximum possible number of simultaneous sources on the road are two vehicles, either passing in opposite directions or one overtaking the other in the same direction.

6.5 Conclusions

This section has presented a derivation of equations to describe the geometrical relationships between a microphone array adjacent to the road and a moving sound source. The equations can be used to model a moving source based on time delay between microphones. Centered on a reference point when inter-microphone time delay is zero, the equations describe the microphone array observation angle and time delay, as well as source velocity and distance travelled. These equations were used when developing a signal analysis method. Using the representative equations, typical situations were modelled for a range of parameters. The modelled scenarios revealed the impact of choosing certain parameter values, such as sampling frequency and inter-microphone distance. This knowledge was used to select system parameter values used when performing experiments.

It was found that knowledge of the distance between two microphones is very important, as this parameter is central to determining the correct model shape for pattern analysis. The distance between the two microphones directly influences the maximum measurable time delay and indirectly determines the maximum number of observations and observed road length. The sampling frequency specifies the measured time precision and hence the location precision and vehicle velocity accuracy.

Based on simulations, a high sampling frequency is necessary to be able to measure velocity to a reasonable accuracy, however this will reduce the processing speed of the system. The choice of parameters are also relevant to satisfy assumptions made in the signal processing approach, for example the distance to the road and audio signal window size affect the far-field and signal stationarity assumptions respectively.

The equations derived in this chapter to model a moving source are central to the proposed traffic monitoring system. The model equations will be used in the following chapter when comparing actual data against simulated behaviour.

Chapter 7

Automatic Vehicle Detection Methods

A time-delay based sound source tracking approach has been described in previous chapters. This method requires a technique to interpret available data in order to extract vehicular characteristics. Therefore, techniques to automatically determine vehicular data based on pattern recognition and extraction are discussed in this chapter. The purpose of these methods is to analyse vehicular data and correctly determine their quantity and behaviour.

There exists a wide variety of applications where pattern recognition is used, reflecting a rich and diverse range of pattern recognition research areas. For example, pattern recognition is used in automated speech recognition, fingerprint identification, iris scanning, optical character recognition, DNA sequence identification, and much more [55, 36, 89, 177].

In some situations, a simple approach that produces tolerable results is more appropriate than a highly accurate and demanding method. For this reason, a vehicle monitoring system based on sound amplitude and frequency spectrum is described in Section 7.1. It is expected to decrease in performance in the presence of noise, but sets an accuracy level from which the other methods may be evaluated. It uses audio signal recordings via a single microphone as the source data. Two other approaches are also discussed that use cross-correlation information as the source data. The first cross-correlation approach filters or sifts the cross-correlation array to extract a data subset containing the most "useful" information. This subset is then used in the decision-making process, as presented in 7.2. Section 7.3 describes the second cross-correlation approach, in which a decision is made based on an integration or combination of all data before determining the best-fitting shape model. The performance of the volume-based method will be evaluated and compared against the cross-correlation pattern extraction approaches in Chapter 8.

7.1 Vehicle monitoring based on sound amplitude and frequency spectrum

This section describes a simple and efficient approach to automatic traffic monitoring, based on the sound energy measured by a single microphone adjacent to the road. The intention is to use the amplitude of the audio signal obtained by a microphone to detect the event of a vehicle passing. The only information that may be determined directly from a single microphone signal are the short and long-term changes in acoustic amplitude. A temporary increase in amplitude indicates a change in the surrounding environment. The microphone is indiscriminate, since it measures all sources of noise arriving at its surface.

7.1.1 Algorithm for vehicle monitoring based on sound amplitude and frequency spectrum

The acoustic amplitude-based event detection process is based on a smoothed, filtered version of the original audio signal, where some processing is implemented to shape the signal into a suitable form for analysis. The steps of the signal processing algorithm are outlined in Table 2, illustrated in Figure 7.1 and described in this section.

It is desired to locate temporary increases or local maxima in the sound amplitude vector. To do so, a 12-second section or temporal window of the audio signal is first isolated. The windowed signal is then smoothed to obtain a general indication of its shape. Smoothing is performed by implementing a series of steps: obtain the absolute version of the signal, retain the maxima of groups of samples and apply a low-pass filter. For each group of samples of length 0.1s, a 10th order low-pass Butterworth filter with a cutoff frequency of 662Hz is applied. Using the smoothed signal, the next step is to locate peaks in the signal. To locate all local maxima and minima, the first derivative of the smoothed signal is obtained. Sign changes indicate local extrema, therefore any instances in the first derivative where the sign changes from + to -, or - to + indicates a local maximum or minimum respectively.

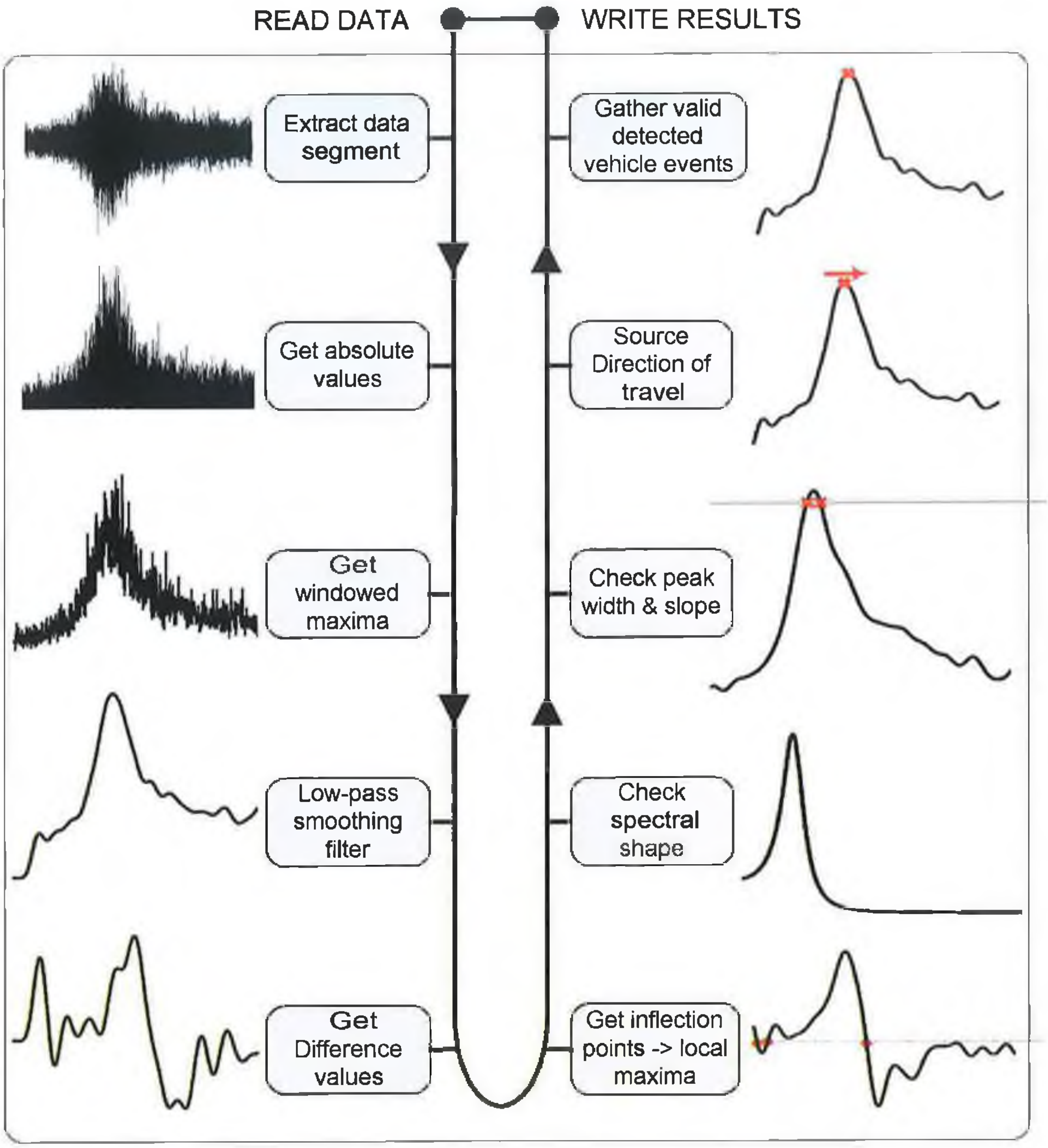


Figure 7.1: Illustration of algorithm steps for vehicle monitoring based on sound amplitude and frequency spectrum

Algorithm 2 Vehicle detection based on sound amplitude and frequency spectrum

1. Isolate a section or window of data;
2. Convert to absolute non-negative values;
3. For every 10 ms of samples obtain the local maximum, thereby reducing the signal size and smoothing the overall shape;
4. Obtain the difference vector to represent the 1st derivative of the signal;
5. The zero-crossing locations of the difference vector indicate the locations of points of inflection. + to - transitions indicate local maxima and - to + transitions indicate local minima;
6. The validity of local maxima as candidate vehicles passing are tested according to the following conditions: if enough time has passed, if sound amplitude is above a minimum threshold;
7. Test whether the frequency spectrum shape is flat and broad. The magnitude of peaks above a certain frequency should not be 25% larger than the previous local minimum;
8. export results and repeat for the next window.

Once a local maximum is found, the next stage of analysis is to determine whether or not the amplitude peak is an indicator of a passing vehicle. A number of criteria are used. The first criterion is whether the magnitude of the peak is above a minimum noise threshold, set at a percentage multiple of the background noise. An adaptive noise amplitude threshold is required that adjusts to long-term changes in background noise. Specifying the time-frame used to define background noise level is a non-trivial task and not directly relevant to the purpose of this research. For this reason, it was decided to obtain the average background noise amplitude from the average sound amplitude over a duration of 10 times the window size. The minimum noise threshold was set at 400% of the background noise amplitude. The same thresholds and parameters were used for all experiments.

The next criterion attempts to distinguish whether individual peaks represent individual vehicles or multiple sounds emitted from the same vehicle. It measures whether sufficient time (0.3s) has passed between two peaks to be considered as individual vehicles, as opposed to individual axles from a single vehicle. The criterion is satisfied if enough time has passed between the peak under evaluation and the previous local minimum.

The final criterion considers the frequency spectrum of the signal around the time of the local maximum. As described in Section 3.3.3, we know that vehicle noise consists of a relatively flat, wide frequency spectrum. Even at low velocities, individual frequency components do not dominate significantly. As a result, the frequency spectrum is generally flat and broad. Therefore, the criterion tests whether the signal frequency spectrum has any significantly protruding frequency components. In order to do so, the frequency spectrum local maxima are evaluated to determine whether (a) spectral peaks occur above a certain frequency and (b) any spectral peak magnitudes above a certain frequency are 25% larger than previous local minimum. This is measured by determining whether the magnitude of peaks above a certain frequency are 25% larger than previous local minimum. If not, then the frequency spectrum is flat enough to be considered a passing vehicle.

Once all the criteria are satisfied for an acoustic amplitude peak, it is considered to represent a passing vehicle. The results are given for a particular window, and the next window of audio data analysed. In order to capture peaks at the edge of the window, successive windows are overlapped by a 1-second interval at each end of the window, resulting in a total window length of 12 seconds. Peaks are then extracted from the non-overlapping central region of 10s duration. The process is repeated for the duration of the entire audio signal.

7.1.2 Analysis of algorithm using sound amplitude

One of the difficulties with an approach using sound amplitude is that it is difficult to identify whether a temporary increase in noise is due to a single noisy vehicle or a group of quiet vehicles in close proximity. A loud truck can acoustically mask successive quiet vehicles long after the truck has passed the microphone. Therefore, estimating the quantity of vehicles present is highly prone to errors.

A second challenge to the method of sound amplitude is background noise. Where there is uncontrollable background noise, the robustness of a system depends on its ability to detect and distinguish the sound of interest from all other sounds that may occur at the same time. In the case of outdoor monitoring of vehicular traffic, there will always be some element of artificial, human or nature-generated background noise. For this reason, a sound amplitude-based traffic monitoring system using a single microphone is certain to fall short of 100% accuracy in all conditions. This is reflected in the experimental results described in Section 8.4.1, where approximately 50% of vehicles are detected based on sound amplitude.

Difficulties arise when it is necessary to specify what is meant by an audio event, since the definition depends on distinguishing characteristics of the desired event in contrast with the audio-based background environment that is to be ignored. The definition of an *audio event* must be specific enough for a system to be capable of returning reliable results, yet be flexible enough to cater for a range of different sounds. For example, an audio-based system may be required that detects a night-time intruder in a building. Indicative noises include alarms, doors or other noise at a time when such noises should not occur. The system can either be designed on the basis of the acoustical properties of a single alarm or the general characteristics of all alarms.

There is a prerequisite of knowledge defining the acoustical properties of an audio event to be detected. Without this knowledge, it is impossible to determine what we want to detect. These properties place boundaries on what type of sound is "relevant" or "irrelevant". While well-designed at the time when the system is developed, such boundaries of relevance must be routinely re-considered over time, in case the event characteristics change. Furthermore, background noises typically change over time. These changes demand an adaptive background filter that is updated at a rate that reflects background noise changes without accidently including temporal foreground noise changes. Background noise that is similar to the audio event will be falsely classified as an event.

Audio event detection based on single or distributed microphones is relevant for multi-sensor security or surveillance applications, where a detected *audio event* can activate technology such as vision-based surveillance. Similarly, an increase in amplitude can serve as an early warning to a traffic monitoring system that traffic is approaching, which in turn activates more sophisticated and power-hungry traffic monitoring systems.

Prior research attempts to use single microphones to monitor traffic have already been discussed in Section 2.1.9 and 4.4.2. There have been no successful published approaches using sound amplitude and a single microphone to directly determine vehicle velocity or location. Although the simplicity and efficiency of using noise amplitude to detect events in the surrounding environment is appealing, it is unlikely to present reliable or accurate results under all conditions.

In conclusion, a traffic monitoring approach based on acoustic amplitude will have difficulties in accurately determining the quantity of vehicles. However, it may indicate the presence of vehicles, even if the number is prone to errors. Chapter 8 describes experimental results based on the sound amplitude-based approach described here.

7.2 Vehicle monitoring by correlation peak tracking

It is desired to develop an approach that efficiently detects and tracks multiple independent sound sources in a cross-correlation array. The method should take into account the cross-correlation array characteristics described in Section 5.4. It is not necessary to analyse complete cross-correlation sequences, only the peaks in each vector that may indicate τ for sound sources present. For this reason, it was decided to develop a pattern extraction method based on detecting, linking and tracking peaks in the cross-correlation data. The objective is to minimise storage memory requirements and maximise speed due to the reduction of cross-correlation data.

The peak tracking approach implemented in this thesis is related to an area of image processing in which edge pixels are detected in an image and contours determined by linking then modelling these edge pixels [121, 184]. Edge detectors yield pixels in an image lying on edges [31, 52, 70, 182], while edge linking attempts to collect these pixels together into a set of edges [11, 122, 129]. An edge linking algorithm yields an ordering of successive edge points based on some predefined criterion functions such as continuity and connectivity. The challenges of edge linking include the fact that small pieces of edges may be missing and small edge segments may appear to be present due to noise where there is no real edge. A problem with this approach is that errors made in edge detection propagate to edge linking without opportunity for correction.

A related approach is also described in speech processing research literature by McAulay [158, 124]. Part of the objective of his work was to locate and track the behaviour of sinewave frequency peaks of a speech model over time. McAulay developed a rule-based algorithm in which different peak trails are not allowed to overlap, split or merge. These constraints make his approach inappropriate for the purpose of linking peaks in the cross-correlation array in this work, since it is necessary to allow time-delay patterns to overlap, split and merge according to the expected event characteristics. A similar approach is developed that is adapted to suit the cross-correlation array characteristics due to vehicular traffic where one key difference from McAulay's algorithm is the ability of the implemented approach to handle crossing paths.

7.2.1 Overview of peak tracking method

The aim of the cross-correlation peak tracking approach is to minimize the volume of data stored and analysed, by tracing the path of salient data and comparing the path behaviour to what is expected of a desired event. It is a bottom-up[1] and reactive method, initiated by prominent peaks in the cross-correlation array. Only the larger peaks in each cross-correlation sequence are selected, the remainder of the array is

[1] There are two methods for developing an algorithm: top-down and bottom-up. The top-down method approaches the problem by starting with the big picture, i.e. a large volume of data, and decomposing it into manageable units. In contrast, a bottom-up method starts with a small selection of information and builds on it to works upward to the top.

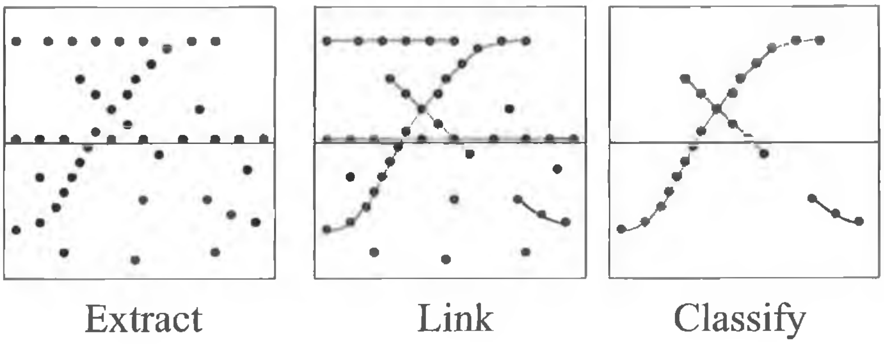


Figure 7.2: Graphical illustration of peak-tracking stages: (a) Extracting time-delay peaks from successive cross-correlation sequences (b) Linking peaks in close proximity with similar behaviour (c) Classifying peak trails according to moving source model parameters

discarded. For successive cross-correlation sequences over time, the propagation of each selected peak is analysed to form peak trails or paths. The resulting paths are analysed with reference to the expected moving source behaviour to produce a list of detected events. No assumption is made regarding the quantity or type of moving sources present in the data, to allow for the presence of multiple simultaneous sources. The method consists of three distinct stages:

1. Prominent peak extraction, described in Section 7.2.2;

2. Link peaks, described in Section 7.2.3;

3. Classify events from peak trails, described in Section 7.2.4.

Figure 7.2 is provided to illustrate the result of each stage. Each stage requires different techniques and distinct approaches to produce satisfactory results.

7.2.2 Extraction of relevant cross-correlation peaks

The output of the prominent peak extraction method are the peaks of particular interest in the cross-correlation array for every time instance. Two windowed audio signals that originate from co-located microphones adjacent to a road, are first cross-correlated. This results in a cross-correlation sequence based on a particular time, as described in Section 5.2. For every sequence, local maxima or peaks with a magnitude above a particular threshold are extracted. The threshold is defined as a percentage of the sequence mean, set at an order of magnitude larger than the sequence mean.

Figure 7.3(a) shows a cross-correlation sequence containing a number of local maxima. Two of the peaks highlighted in red exceed the threshold, and are therefore extracted from the sequence. A series of characteristic values are stored for every peak retained. These are illustrated in Figure 7.3(b). The labels are described in detail in Table 7.1.

A naïve peak extraction approach would be to retain only the dominant peak and track this value over time. However, this is not feasible for several reasons. Firstly this assumption may work for noise-free signals of sequential vehicles in a single lane, but not for multiple lanes or bidirectional traffic. Secondly, as described in Section 5.4, when the vehicle is in close proximity to the microphone array, the source may be observed as two sources or peaks as opposed to just one. Therefore simply tracking the dominant peak is insufficient. It is necessary to be able to detect, track and evaluate multiple candidate peaks as well as their behaviour over time. Instead of tracking the predominant peak, *all* the peaks in the cross-correlation sequence above a threshold are tracked. This is to ensure that in case of multiple sources, or false indication of the predominant peak, a correct data interpretation is still possible. By tracking the peak behaviours over time, it soon becomes apparent which peaks represent moving sources such as vehicles, and whether there is a single source or there are multiple sources.

The process is repeated for each successive cross-correlation sequence, generating

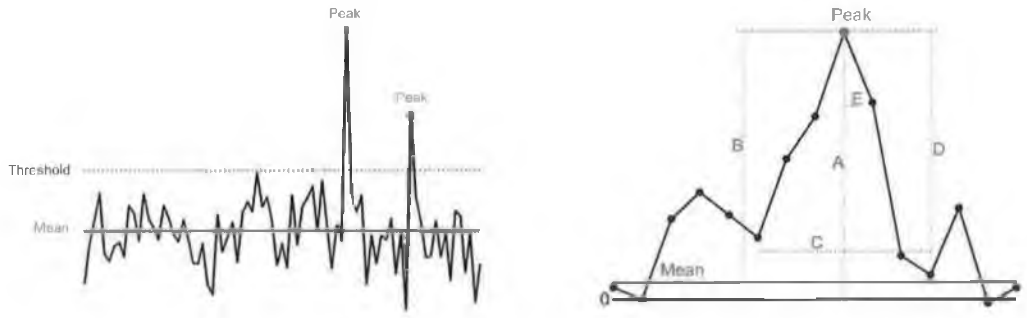


Figure 7.3: (a) Picking peaks in the cross-correlation sequence for tracking (b) Characteristic parameters of each extracted peak, described in Table 7.1

Table 7.1: Peak parameters stored during the extraction of relevant cross-correlation peaks

A	peak magnitude relative to $y = 0$
B	peak magnitude as a percentage of sequence average
C	base width of peak between nearest local minima
D	average height of peak relative to nearest local minima
E	average peak roll-off to nearest adjacent points

more and more isolated peaks. This is shown in the extraction stage of Figure 7.2.

7.2.3 Linking of peaks to create events

The extracted peaks need to be compared and any existing patterns determined, in order to make decisions on moving source present in the data. Any related peaks are linked after new peaks have been extracted from the latest cross-correlation sequence. This section describes the implemented method in which the latest or *new* peaks are compared to existing peak *trails*, then matched if appropriate.

For every new cross-correlation sequence, N new peaks are compared to M existing trails. N and M are positive whole numbers that do not have to be equal. Rules describing acceptable trail behaviour are defined as follows:

1. Each trail may be born, die, or sleep before being reborn;
2. Trails may cross each other in different directions;
3. A single trail may split into two;
4. Two trails may merge into a single trail;
5. Successive points may not exceed a specified maximum distance beyond the previous point on the same trail;
6. To be consistent in detecting vehicle behaviour, individual trails may not change direction by more than 45 degrees at a time;
7. Trails may become inactive for a limited time.

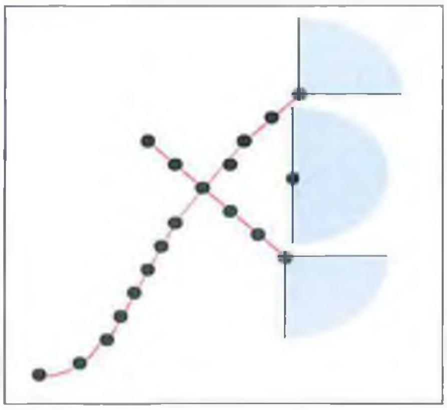


Figure 7.4: Trail target ranges for future peaks

Each live trail has a target zone for new peaks, highlighted in grey in Figure 7.4. The target zone is governed by the overall trail direction. The zone size and shape is constrained by the previous rules describing trail behaviour. For every cross-correlation sequence generating another round of peaks, a combination of the following actions are taken:

1. If no new peak is detected within a target zone and the respective trail has been inactive for a number of successive windows, it is considered to be dead;
2. If no new peak is detected within a target zone, the respective trail is inactive for this round;
3. If a single new peak is detected within a single target zone, the trail is updated to include it;
4. If a new peak falls outside all target zones, it causes the birth of a new trail. Similar to all existing trails, its survival depends on the presence of future points within its target zone;
5. If multiple new peaks are detected within a target zone, or a peak falls within multiple target zones, a linking decision is made to determine the course of action.

A simple method for linking multiple spectral peaks was described by McAulay in [158, 124]. He derived a sinusoidal model for a speech waveform and obtained the sinewave frequency peaks by locating the peaks in the Fourier transform of the speech signal for each frame. The spectral peaks are tracked over successive frames according to a series of ad-hoc rules. In McAulay's method, different peak trails are not allowed to overlap, split or merge. These constraints make his approach inappropriate for the purpose of linking peaks in the cross-correlation array in this work, since it is necessary to allow time-delay patterns to overlap, split and merge according to the expected event characteristics. A *dynamic programming* approach is used in this thesis to accurately resolve the problem of correctly linking multiple new peaks described in point 5 above. Developed by Richard Bellman [19], dynamic programming is a popular optimization method based on the *principle of optimality* [153, 154]. The objective of this principle is to determine the cost of each possible decision, then make the best set of choices by minimizing the overall cost.

The dynamic programming algorithm used for linking N new peaks to M existing trails has three consecutive stages:

Build cost space The cost space describes the cost of all linking combinations between the new peaks and existing trails. The building of the cost space is described in Algorithm 3.

Determine best path The best path through the cost space is established using backward propagation by considering the path of all linking options, to form a series of back pointers. This is repeated to form the second-best path and so on, as described in Algorithm 4.

Link peaks and trails For each peak, the best backpointer to a particular trail is selected. Assuming a match has been found, the best-matching peak and trail are linked together. Linking peaks and trails is described in Algorithm 5.

The meaning of variable names in Algorithms 3, 4, 5, are as follows:

p(N) is the list of N new peaks extracted from the latest cross-correlation.

tr(M) is the list of M trails.

C(i, j) is the K x K total path cost.

K is the number of all possible τ values in the cross-correlation array.

dirWeight is the absolute difference between the trail direction and new peak location.

magWeight is the absolute difference between the last trail peak magnitude and the new peak magnitude.

pathCost stores the costs of a particular path, while pointer describes the paths through the cost matrix.

pointer describes the paths through the cost matrix, while pathCost stores the cost of these paths.

candidateCost used to test against other path costs, updated when a better value is found.

bestCost is the lowest possible overall cost of linking a peak.

bestBackPointer is a pointer to the best path.

Algorithm 3 Dynamic Programming - Build cost space

```
initialise the cost matrix to infinity; $C(i,\ j) = +\infty$
for $i = 1$ to $i = N$ do
  for $j = 1$ to $j = M$ do
    if $|i - j| < maxDist$ then
      $dirWeight = |tr(j)\ direction - p(i)\ location|$
      $magWeight = |p(i)\ magnitude - tr(j)\ magnitude|$
      $C(a,\ b) = dirWeight + \frac{magWeight}{4}$
    end if
  end for
end for
```

7.2.4 Classification of peak trails

Once the new peaks and existing trails have been appropriately linked, the next step is to examine the trails for possible event classification. Each trail is analysed, including trails that have recently become inactive during the latest linking iteration. To consider the trail as a passing vehicle, it must fulfill the following criteria:

1. The trail lifetime must be sufficiently long to return reasonably accurate parameters when examined;

2. The trail must span a minimum quantity of different τ values.

Algorithm 4 Dynamic Programming - Determine best path

```
for $i = 1$ to $i = N$ do
    for $j = 1$ to $j = M$ do
        initialise bestBackPointer to -1
        for $p = 1$ to $p = M$ do
            $candidateCost = pathCost(i - 1,\ p)$
            if $candidateCost < bestCost$ then
                $bestCost = candidateCost$
                $bestBackPointer = p$
            end if
        end for
        $pathCost(i,\ j) = bestCost + C(i,\ j)$
        $pointer(i,\ j) = bestBackPointer$
    end for
end for
```

Algorithm 5 Dynamic Programming - Link peaks and trails

```
initialise bestBackPointer to -1
initialise i to N
while $i > 0$ do
    for $j = 1$ to $j = N$ do
        $candidateCost = pathCost(i,\ j)$
        if $candidateCost < bestCost$ then
            $bestCost = candidateCost$
            $bestBackPointer = pointer(i,\ j)$
        end if
    end for
    if a match has been made then
        $bestBackPointer = pointer(i,\ bestBackPointer)$;
        match trail(bestBackPointer) to peak(i);
    else
        create new trail from unmatched peak
    end if
end while
```

If both criteria are satisfied, then the trail is iteratively compared against versions of the derived mathematical model to return optimum matching values for v and t_{ref}, as illustrated in Figure 7.5(a).

A *least squares* process is used to find the matching model parameters for each trail. Least squares is a mathematical optimization technique that attempts to find a function which closely approximates a series of measured data, often termed a *best fit*. It attempts to minimize the sum of the squares of the offsets between points generated by the function and corresponding points in the data. Consider Figure 7.5(a) showing data and a corresponding model. Suppose that the trail data set consists of the points $(x(i),\ y(i))$ with $i = 1, 2, \cdots, N$. It is desired to find the parameters v and t_{Ref} of the analytical model τ_m derived in Section 6.1.4, such that $\tau_m(i) \approx y(i)$ and the parameter values minimize the sum of the squares of the offsets. A single residual value r is obtained for each unique set of parameter values, by summing the square offsets of the data from the model function. Illustrated in Figure 7.5(b), the vertical offsets from a function are minimized as opposed to the perpendicular offsets for practicality. This is analytically denoted as:

$$r = \sum_{i=1}^{N} (d(i) - \tau_m(i))^2, \tag{7.1}$$

and graphically illustrated in Figure 7.5(b). The equation is used iteratively to obtain the residual value for all values of the model parameters v and t_{ref}. If every possible parameter value in the range of velocities and reference times was used to calculate a residual, the complete 2D search space would be built by brute force. A faster and more efficient approach can be used to converge rapidly on the model parameters that returns the smallest residual value. The recurring approach consists of two stages: optimization of t_{Ref} and optimization of v. During the first stage,

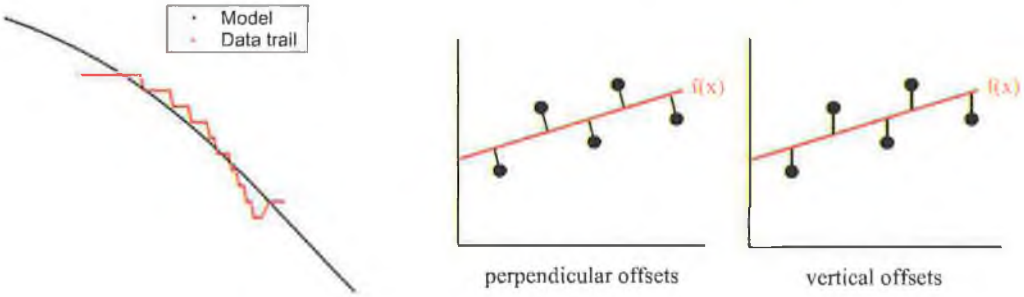


Figure 7.5: (a) Matching model to trail (b) Least squares offsets may be obtained

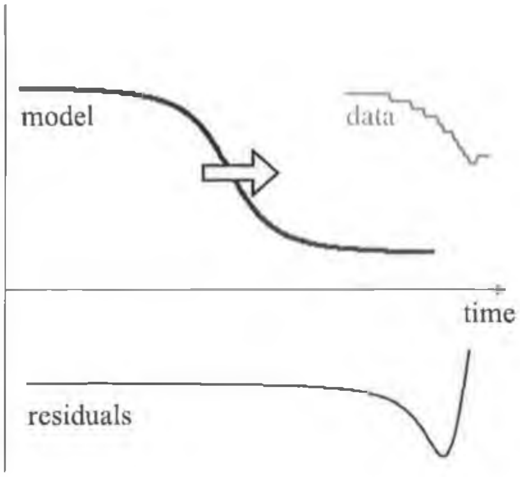


Figure 7.6: Sliding model over data to miminize residual

the optimization of t_{Ref} is performed while v is kept constant. For a particular velocity, the model function is iteratively compared to the data trail for increasing t_{Ref} values. This effectively slides the model over the data along the time x-axis, as shown in Figure 7.6. For each time instance, a residual value is obtained. These are combined to form a time-based residual vector. The minimum residual value reflects the optimum t_{Ref} value for that particular velocity, and is stored for comparison.

The second stage, optimizing v, is performed by testing three different velocities; two from the outer limits of the range of possible velocities (v_{min} and v_{max}) and one from the midpoint (v_{mid}). The three models with these velocities are individually tested to optimize t_{Ref} according to the previous step. Analysis of these three models returns three residual values. The two smallest residuals from the group of three determine the outer limits for the next iteration, thereby updating v_{min}, v_{mid} and v_{max}.

Every time these two steps are repeated, the search area is halved. Over numerous iterations, there is a convergence on optimum v and t_{Ref} parameter values that return the smallest overall residual. The iterations continue until one of the following criteria is satisfied:

- the residual value is sufficiently small and the process has converged;
- the number of iterations reach a maximum amount (to prevent infinite repetition without convergence);
- the residual value is not being reduced by further iterations (to prevent infinite

repetition without convergence).

The parameter values of the optimum model function are then assumed to closely approximate the data in that particular trail. These values are added to the output results. The classification process is repeated for every recently dead or inactive trail. Duplicate parameters from multiple trails arising from the same moving source are eliminated by post-processing the output results for comparison.

7.3 Correlation-based vehicle monitoring based on shape matching

The previous approach emphasised the time sequence of correlation vectors and used a process of peak identification as a data reduction step before a semantic "line-tracking" step. That approach is reminiscent of line tracking following edge-detection in computer vision, and points towards consideration of the time sequence of correlation vectors as a 2-D data array or image.

Inspired by the Generalized Hough Transform, the second cross-correlation approach searches for regions or shapes of high correlation in the cross-correlation array that match the time-delay shape model of a passing vehicle. All array values within the region of a particular shape model contribute in making a decision regarding that particular model, in a similar manner to Hough parameterised line/curve detection. This is repeated for all model parameter values, with the results being mapped into the model parameter space. Described in Section 7.3.1, points or regions of high magnitude in the parameter space indicate a high correspondence between a model (and hence a passing vehicle), and the data. In this manner the detection of a vehicle is robust to some noise in the cross-correlation array, since many different values contribute to the overall decision. The approach integrates "votes" for a particular model rather than being based on noise-emphasising derivatives, which could be claimed of the previous model.

7.3.1 Overview of the Hough transform

The Hough transform is a global, robust technique for the detection of predefined shapes in data [79, 105]. It was first introduced by Paul Hough [144] in 1962 for identifying the slope of lines in an image[2]. The Hough transform can be used successfully for the detection of overlapping or semi-occluded objects in noisy data, and is an important technique in applied Computer Vision. There are two widely used types of Hough transforms: the classical Hough transform, which is used for straight line detection, and the Generalized Hough Transform, which is used for arbitrary shapes.

Line Hough transform

Mathematically, the Hough transform is simply an *integral transform* in that it integrates all values in the image $I(x,\ y)$ along the shape of interest [104]. The Hough transform for a line $H(p,\ \theta)$ in polar co-ordinate notation is defined as:

$$H(p,\ \theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x,\ y)\delta(p - x\cos\theta - y\sin\theta)dxdy. \tag{7.2}$$

The delta function has the effect of sampling $I(x,\ y)$ where the delta function's argument is zero, which is along the line $p = x\cos\theta - y\sin\theta$. Moreover the delta function has the effect of forcing the integral to ignore all other points in $I(x,\ y)$.

A line can be represented using the *slope-intercept* form

$$y = mx + c, \tag{7.3}$$

where m is the slope and c is the intercept. However, this equation is not stable when m and c approach infinity. Duda and Hart [56] proposed a more appropriate representation of a line in the *normal form*:

$$p = x\cos\theta + y\sin\theta. \tag{7.4}$$

This equation specifies a line passing through Cartesian co-ordinates $(x,\ y)$ that is perpendicular to the line drawn from the origin to $(p,\ \theta)$ in polar space, as illustrated in Figure 7.7(a). For each point (x, y) on such a line, $(p,\ \theta)$ are constant. The normal form proves to be better than the slope-intercept form, as it is numerically stable for matching lines of any angle. Therefore, the normal form representation is used in this thesis.

[2] Deans [48] later showed that the Hough transform is a special case of the pre-existing Radon transform [160].

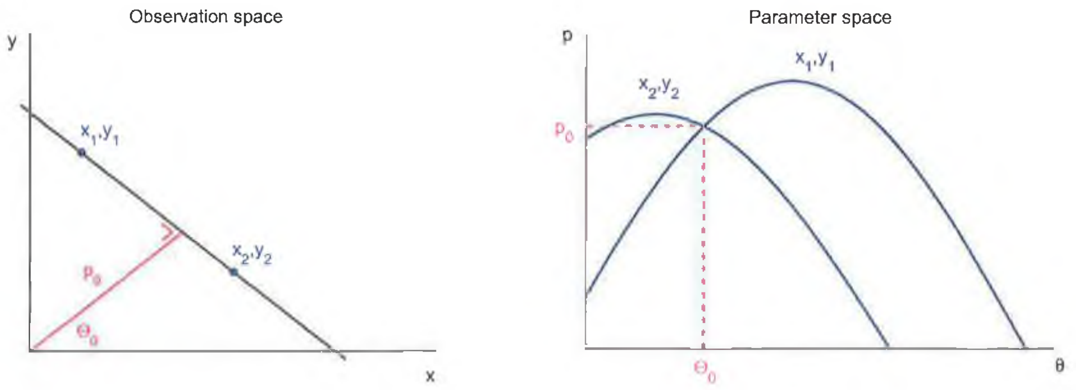


Figure 7.7: (a) Observation space and (b) Parameter space for the normal parametrization of a line where $p = x \cos\theta + y \sin\theta$

To understand the Hough Transform, it is important to mention *parameter space.* The parameter space is a representation of values for which the axes are the parameters of an equation. In the area of Hough transforms, the parameter space is also called a *Hough space.* There are as many dimensions in the parameter space as the number of variables in the equation describing the shape of interest. For example a line may be described with two variables; the slope and intercept. The parameter space of a line is a two-dimensional plane, where each axis represents the slope and intercept values respectively. For a Hough transform designed to detect lines, each point in the parameter or Hough space $H(\theta,\ p)$ corresponds to a line at angle θ and distance p in the original space. For each point in the original space, sometimes called the *observation space*, all the lines which go through that point at a discrete set of angles are considered.

Consider now Figure 7.7(a). The observation space axes represent the $(x,\ y)$ coordinates. To obtain the correct values in the parameter space shown in Figure 7.7(b), all possible values of p and θ are considered using Equation 7.4. Consider now two points $(x_1,\ y_1)$ and $(x_2,\ y_2)$ located on the line illustrated in Figure 7.7(b). They translate to individual curves in the parameter space in Figure 7.7(b), whose axes represent θ and p respectively (plotted in solid lines). The point of overlap of the two curves $(\theta_0,\ p_0)$, indicates the parameter values of a line that intersects both points in the observation space.

Discretising the Hough space results in an array of bins, called an *accumulator array.* For each coordinate calculated, the accumulator array is incremented by the value at the corresponding $(x,\ y)$ location in the image. The increment to a bin of the accumulator array is sometimes referred to as a *vote*. After considering all the lines through all the points, a high-valued accumulator or global maximum in the Hough space indicates the presence of a line. In the case of multiple global maxima, it is possible that there are multiple lines present in the original image, such as the case where the simple source image (Figure 7.8(a)) results in two maxima in the Hough space (Figure 7.8(b)). Therefore, once the Hough space has been created, there is an additional step of selecting and interpreting clusters to determine the number of detected shapes, as well as shape parameters.

Generalized Hough Transform

Generalized Hough Transforms, developed by Ballard [16], extract the shape in its entirety rather than decomposing the image into its component features (such as lines). It was introduced to enable dealing with shapes which cannot be represented analytically.

The Generalized Hough transform is an extension of the Hough transform for lines applied to other shapes of arbitrary complexity. In the case of shapes that are not easily expressed using a small set of parameters, the points on the shape can be explicitly listed by creating a look-up table that contains all of the $(x,\ y)$ coordinates for the target shape. The generalized Hough transform is particularly useful for detecting 2D object shapes with specific orientations and scales.

It is not necessary that the curves detected by the Hough transform be described in a parametric equation. The Hough transform can be generalised into a voting algorithm that implements template matching efficiently. Template matching is where a replica of an object is compared to all unknown objects in the image field. If the template match is sufficiently close, the unknown object is labeled as the template object.

Algorithm 6 encodes the shape of the object boundary in a table for efficient access. One point on the object is chosen as the reference point. By definition, the location of the reference point in the image is the location of the object. For each image gradient point at (x, y) with gradient angle θ, the possible locations of the reference point are given by ... Each possible reference point location is incremented. The location of the peak in the parameter space is the estimate for the location of the

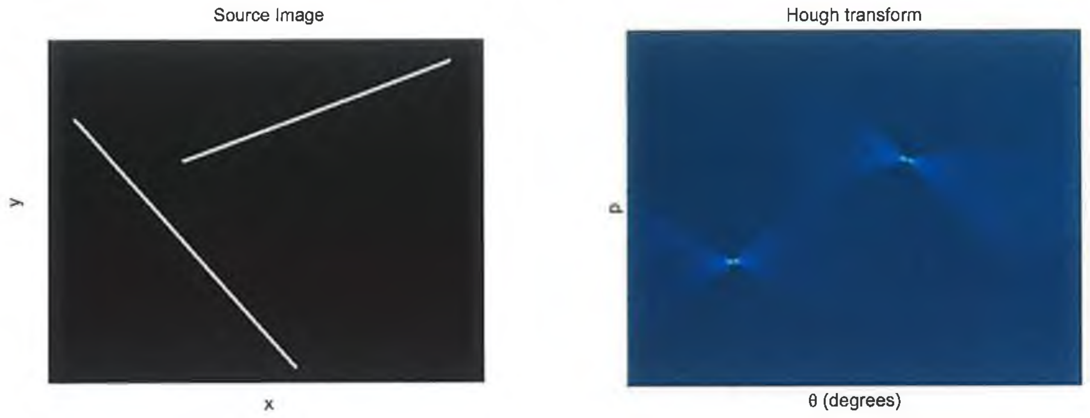


Figure 7.8: (a) Source image (b) Hough transform

Algorithm 6 Voting algorithm for template matching based on the Hough transform

1. Pick a reference point on the object;
2. Compute the gradient angles θ_i along the object boundary;
3. For each gradient point θ_i, store the distance r_i and angle θ_i from the reference point.

object. It is not easy to generalize this technique to incorporate changes in scales or rotation [82].

Properties of Hough transforms

The Hough transform technique is rather robust, even when there is a high percentage of gross errors or noise in the data. There is no requirement for points to be connected, or even nearby, for them to vote for the same parameter space location that describes the shape. The Hough transform is a useful method, particularly when the shape is easily expressed using a small set of parameters. On the other hand, a Hough transform is potentially a computationally expensive approach that increases dramatically in complexity with increasing number of parameters. However, since the process of parametric transformation does not make explicit any information concerning connectivity, it may break down when exposed to data containing correlated noise, due to the accidental grouping of data points. This may give misleading results as well as the case when two shapes happen to be aligned. If the amount of data points is not sufficiently large, the maximum peak in the Hough space is not much higher than other peaks. For this reason, the Hough transform is better suited to problems with sufficient data to support the expected result.

To obtain accurate Hough Transforms, the appropriate sampling intervals must be chosen for the parameters. The granularity with which the parameter space is discretised determines how accurately the sought-after target may be positioned. When the bin sizes are chosen too fine, results from a single shape can be placed in different adjacent bins. This causes the anticipated global maximum to be lower in magnitude than expected, due to the contributory votes being distributed across different locations. On the other hand, when the quantization is too coarse, votes from distinct shapes which are close together will lie in the same bin. If the "true" parameters of a shape happen to lie close to a boundary in the quantized parameter space, the votes will be spread over two or more bins, therefore observing single bins may not reveal the peak. Furthermore, because of quantization errors and noise in the measurements, the expected peak in the accumulator may be blurred so that it is not easily detected.

7.3.2 Hough-based approaches

Two different Hough-based approaches to detecting patterns in the cross-correlation array are now described. Both approaches are based on a Generalized Hough Transform, using reference tables. The first approach searches for rectangular regions of high correlation in the data. In this manner the strong presence of lines at particular locations and with a certain range of slopes indicate a moving source. The second method searches for shapes matching the moving source model derived in Chapter 6. This results in the model-based Hough transform detecting the exact model shape, rather than approximating the model with a line. In both approaches, a version of the shape being sought is iteratively applied for increasing parameter values; slope and time of passing respectively.

Rectangular Hough-based approach

The rectangular Hough transform approach is a variant of Hough line detection. A block or rectangular shape is sought in the cross-correlation array. Instead of using only boundary or edge points, all points contained within the shape region are utilized. Since a rectangle is a line with a particular width, the Hough line detection method, described in Section 7.3.1, can be adapted for the purpose of identifying a rectangle. Also, since the expected shape location is constrained by the data characteristics described in Section 5.4, there is no need to explore all outer extremities in the matrix.

The parameter space of the rectangular shape consists of two axes representing the reference time t_{ref} and line slope θ. In the observation space, the rectangle is pivoted relative to the horizontal axis (where $\tau = 0$) for a range of θ values. All cross-correlation values within the rectangle are summed to obtain a single global measure for those particular parameter values. The summed cross-correlation measure is stored at the appropriate location in the parameter space. This is repeated for successive reference times t_{ref} where the block width is defined by the variable w and the height is defined as $2h$. Only data along the central section of the moving source pattern in the cross-correlation array can be considered, as the moving source pattern changes in an increasingly nonlinear fashion for larger values of τ. Figure 7.9(a) illustrates the rectangular block used while Figure 7.9(b) displays the parameter space when a single event is present in the source array. A region of high magnitude is evident in the parameter space as a dark red area.

A two-step approach is adopted to extract relevant information (passage, direction and velocity) corresponding to the detection of a vehicle. First, a subset of time delays are summed as in Equation 7.5, and the moving average $a(t)$ tracked.

$$a(t) = \sum_{\tau=-\tau_{max}/4}^{\tau_{max}/4} R(\tau, t). \tag{7.5}$$

In the event of a significant magnitude change, i.e. when $a(t)$ increases sharply, the system is alerted to the passage of a vehicle. A rectangle of the form shown in Figure 7.9(a) is then applied. The aggregate magnitude AM_i of the cross-correlation matrix R is accumulated over a rectangle of a particular slope, m_i corresponding to angle θ_i for $i = 1, ..., N$.

$$AM_i = \sum_{x=t_0}^{t_{max}} \sum_{y=m_i x - \frac{w}{2}}^{m_i x + \frac{w}{2}} R(x,\ y) \quad i = 1, ..., N, \tag{7.6}$$

where t_0 is the starting time instance up to a horizontal width of t_{max}. Care must be taken to choose an appropriate rectangle height that is not too small, reducing the effectiveness of the directional filter. Rectangle height is determined as a percentage of τ_{max}. By repeating for all possible parameter values, the Hough space is formed.

The next step is to reduce the 2D Hough space to a time-based vector representing the maximum value for each range of values of θ. Figure 7.10(a) shows two one-dimensional sequences representing negative and positive slopes respectively. The 2-D array is the cross-correlation data of two vehicles passing in opposite directions, the two rectangles being superimposed for illustrative purposes. An event is detected by locating maximum values in either negative or positive sequences, where the maximum value, AM_{max} is given by

$$AM_{max} = \max_{m_i i=1,...,N}(AM_i). \tag{7.7}$$

By determining which slope generates the maximum value, AM_{max}, the vehicle velocity v can be estimated. The algorithm to extract events from the parameter space is described in detail in Table 7.

A disadvantage of the described approach is that the rectangular shape being sought is not the same as the modelled shape of a moving source. This means that even a perfect match based on a rectangle does not optimally represent actual source

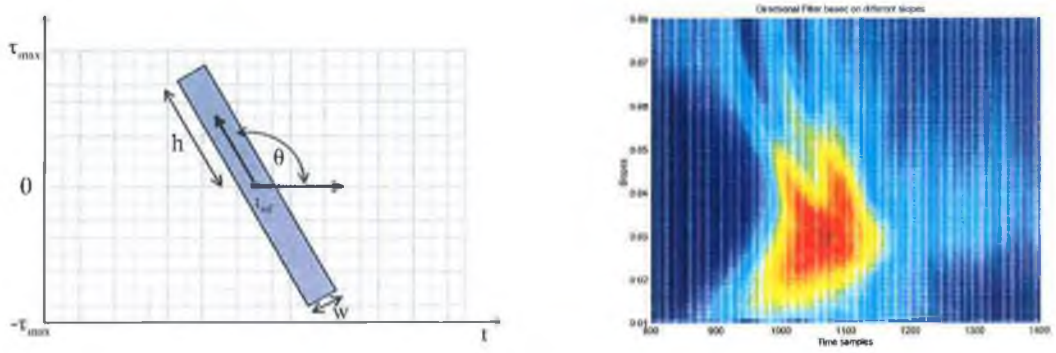


Figure 7.9: (a) Parameterized rectangular shape applied to cross-correlation array (b) Parameter space representing t_{ref} and θ rectangular parameters for a single event

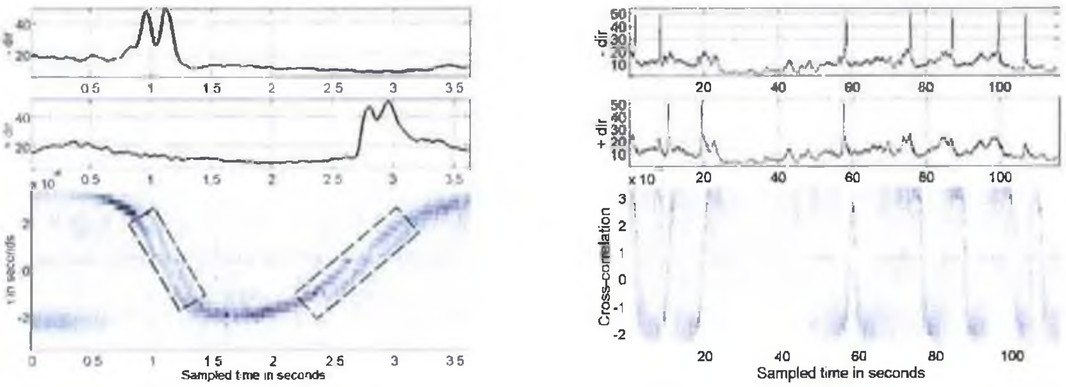


Figure 7.10: (a) Directional filter applied in positive and negative direction to the cross-correlation array of two passing vehicles (b) Positive and negative directional filter applied to two minutes of data with 10 passing vehicles

Algorithm 7 Algorithm to interpret rectangular block parameter space and obtain peaks

1. Create two sequences from the 2D parameter space by retaining the maximum value for every series of velocities at each t_{Ref} in the first vector, and the value of velocity where the maximum value occurred in the second vector;
2. Find the peaks and associated velocities in the magnitude vector based on the zero-crossing of the first derivative, peak magnitude and roll-off. Two thresholds are imposed at this point;
3. For every peak detected, determine the sign and reference time. From this, the direction and time of occurrence of a passing sound source is known. Similarly, the associated source velocity is also known;
4. Present a list of the required traffic data to the user, to include time of occurrence, direction and speed. If necessary, the audio signal around the appropriate times can be analysed to obtain a spectral estimate and used for vehicle classification.

behaviour. Using a more precise shape should not only improve detection accuracy, but also create greater insight into model parameter values such as source velocity. For this reason a Hough-based approach is developed using the derived moving source model.

Model-based Hough Transform

The method described in this section is based on the derived moving source model depicted in Figure 7.11. Similarly to the previous rectangular-based method, the cross-correlation data is mapped into the model parameter space, which is appropriately quantized. Figure 7.12(a) displays the cross-correlation data array for a single passing vehicle and the corresponding parameter space. The details of building the parameter space is described shortly.

Having a completed parameter space does not achieve the goal of automatically detecting moving sound sources. The parameter space highlights the presence of particular shapes in the data. However, the parameter space must be analysed in turn, in order to decide on the number and type of events it represents. For example, the clusters of high values in the parameter space in Figure 7.12(b) indicate possible

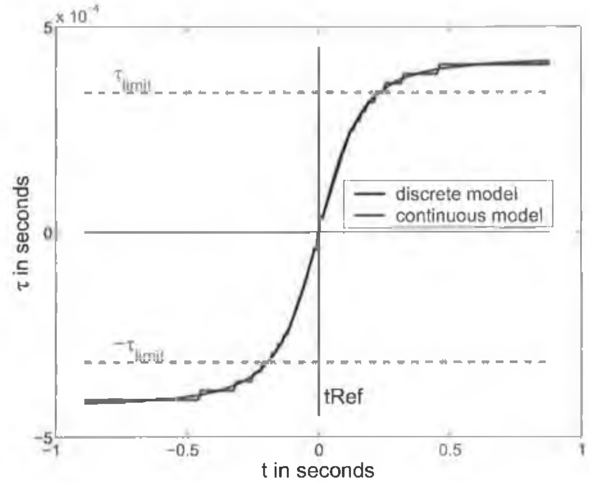


Figure 7.11: Illustration of a discrete and continuous moving source model

events. Therefore it is necessary to continue the process of pattern analysis to reach the goal of automatic traffic monitoring. Clustering techniques are used to determine events and the most likely parameter values in the Hough space. The method used to translate the parameter space data to an automatic list of traffic characteristics is described after the following section.

Building the parameter space

To fill the parameter space with confidence values, each model instance is repeatedly applied to the observation space at every time instance. Therefore the first step is to match the model and observation space. However, the observation space consists of a cross-correlation data array obtained from sampled audio signals with fixed window. This results in quantized values stored in different bins that are dictated by the sampling frequency and window size. The parameterized model is based on a continuous function that does not necessarily match the data bins exactly. Therefore the model must either be quantized in a manner to force it to match the data, or a weighted interpolation of data bins used to more closely approach the model.

Once the model and observation space are matched in size, the model values are used as indices to select data at specific locations in the cross-correlation array. The series of specific locations are positioned along the trajectory of the model as if the model is superimposed on the data array. The values at these locations are averaged, resulting in a single value for that model with the given parameter values. This is

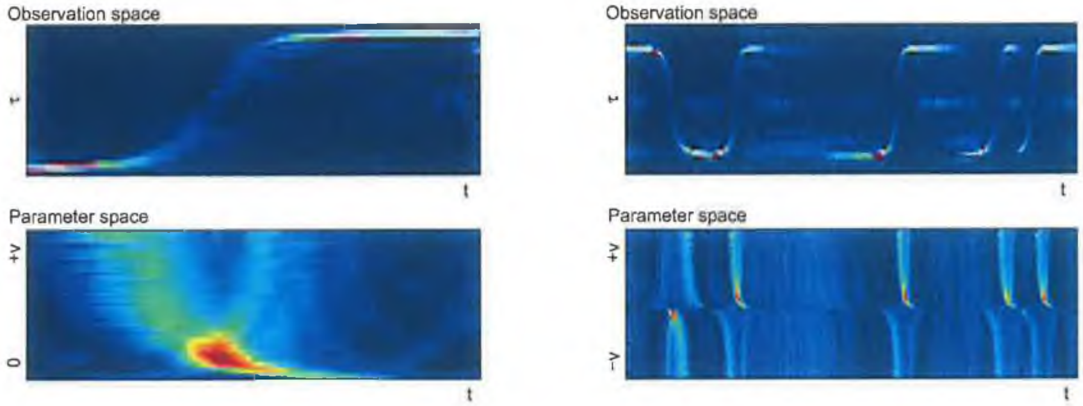


Figure 7.12: Model-based observation and parameter space for (a) a single passing vehicle (b) multiple vehicles in both directions

then stored in the parameter space location that matches the current parameter values. This process is repeated until the parameter space is filled with values. This sequence of steps to build the parameter space $p(t,\ v)$ for all values of t and v is given as Algorithm 8.

Consider Equation 7.8 defining the analytical model for time delay of a source moving at a constant velocity v. It is illustrated in Figure 7.11. The modelled inter-microphone time delay $\tau_m(v,\ t)$ is a function of time t relative to the central reference time t_{Ref} where $t = 0$. The other variables are constants dictated by the microphone geometry and physical environment.

$$\tau_m(v,\ t) = \frac{D}{c}\left(\sqrt{1+\left[\frac{m}{2D}+\frac{vt}{D}\left(\frac{m^2+4D^2}{m^2+4D}\right)\right]^2} - \sqrt{1+\left[\frac{m}{2D}-\frac{vt}{D}\left(\frac{m^2+4D^2}{m^2+4D}\right)\right]^2}\right). \tag{7.8}$$

The values of the model beyond $-\tau_{limit}$ and τ_{limit} in Figure 7.11 describe the moving source position when far away. These values provide little information about the vehicle behaviour as it passes the microphone. Therefore only the portion of the model within the boundaries of τ_{limit} and $-\tau_{limit}$ is utilized to analyse the observation space. Using the model $\tau_m(v,\ t)$ within the boundaries as indices to access the cross-correlation array $r(t,\ \tau)$, the parameter space bin $p(t,\ v)$ is defined as:

$$p(t,\ v) = \frac{1}{N}\sum_{t=1}^{N} r(t,\ \tau_m(v,t)). \tag{7.9}$$

Algorithm 8 Shape Detection - build parameter space

```
Get $r(t,\ \tau_m)$
for i = 1 to i = N do
    for j=1 to j=M do
        Calculate $\tau_m$ for parameters $t_{Ref}(i)$ and $v(j)$
        Quantize $\tau_m$ based on $f_s$ and audio window size
        Isolate the section of $\tau_m$ within $\tau_{min}$ and $\tau_{max}$
        Use the section of $\tau_m$ as indices to access $r(t,\ \tau_m)$
        Average the accessed values and store in p(i, j)
    end for
end for
```

$r(t,\ \tau_m)$ is the cross-correlation array.

t_{Ref} is the reference time parameter, used in the moving source model.

v is the velocity parameter, used in the moving source model.

N is the size of the range of values of t_{Ref}.

M is the size of the range of values of v.

τ_m is the moving source model with parameters t_{Ref} and v.

τ_{min}, τ_{max} is the horizontal boundaries of useful bit in cross-correlation array.

p(i, j) is the parameter space bin at location (i, j).

N is the number of samples in the model, and $r(t,\ \tau_m(v,\ t))$ is a cross-correlation value at the location in the observation space indicated by t and $\tau_m(v,\ t)$. Now that the parameter space is complete, the final step is to interpret it in a manner that moving source events and their parameters are detected.

Parameter Space Interpretation

If the presence of only one shape is expected and the quantization level is suitable, parameter space analysis is a simple matter of finding the global maximum. In reality, multiple shapes may be present, causing multiple local maxima in the parameter space. Additionally, there are a few different situations that complicate an understanding of the parameter space. Firstly, inappropriate quantization of the parameter space can cause evidence of a shape being distributed among multiple parameter bins. This must not however be misinterpreted as distinct shapes. Secondly, relative maxima with few votes are typically not real matches. Finally, evidence from two distinct shapes in close proximity may combine to give the illusion of a single shape with different parameter values being present. In short, the correct interpretation of the parameter space is the key to successfully utilizing the Hough transform-based approach.

Algorithm 9 Algorithm to interpret the parameter space

1. Reduce the parameter space to a single dimension, with a series of values for each direction consisting of the maximum value for every time instance and the original location (i.e. velocity) of the maximum value, shown in Figure 7.13;
2. Determine all the peaks in the maximum value for every time instance above a pre-defined threshold. If a peak occurs in both directions then two simultaneous vehicles are passing;
3. Obtain a subset of the peaks, represented as a red dot in Figure 7.13, that satisfy the following two requirements: peaks must be a minimum distance apart and the local minimum between peaks must be a number of times smaller in magnitude than the peak.

Algorithm 9 was used to interpret the parameter space. It requires the use of three thresholds to find local maxima. A more elegant technique to interpret the parameter space is intended as future research. However, experiments in Chapter 8.4.2

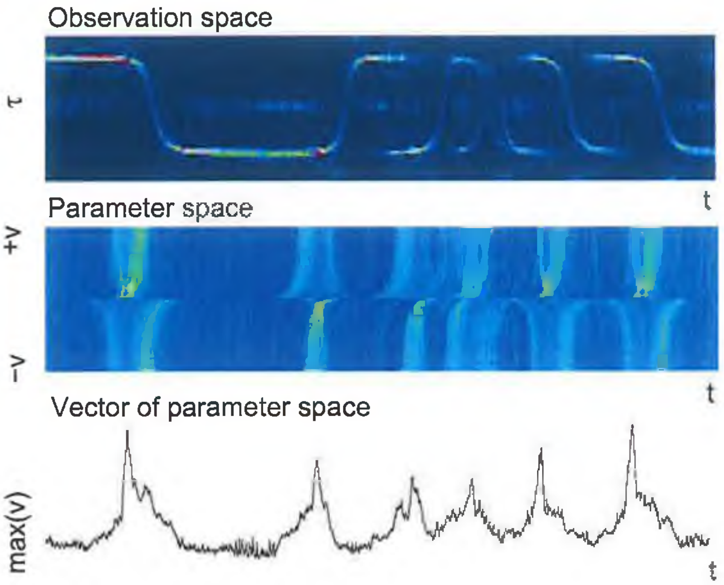


Figure 7.13: Model-based Hough Transform

demonstrate the high accuracy in vehicle detection using this algorithm, therefore it was deemed suitable as a first solution.

7.4 Vehicle Axle detection

From observing the cross-correlation array in Figure 7.14, two distinct moving source patterns can be seen, particularly when in close proximity to the microphone array. These represent noise emanating from the front and rear of the moving source respectively. When moving at higher velocities, the dominant vehicle sound is due to tyre/road interaction; the front and rear axles both contribute sound. For a short while, as the vehicle is close enough to the microphones, it is possible to measure the sounds from these axles as two distinct sources. This is visible in the cross-correlation array as a single source in the distance sometimes becoming two distinct sources when passing by the microphone array. The ability to detect axles based on a cross-correlation approach was described by Chen [39]. However, the author is not aware of publications describing successful implementation of automatic axle detection.

A very brief drop in sound amplitude was heard in the audio signals when a vehicle passes the microphone array. This was particularly true for larger vehicles. It is

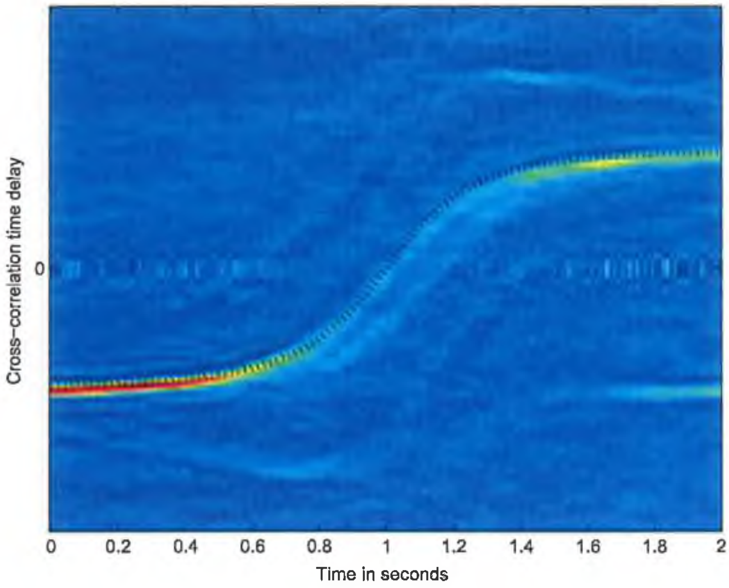


Figure 7.14: Cross-correlation array with model superimposed - to show how model can be used to detect vehicles

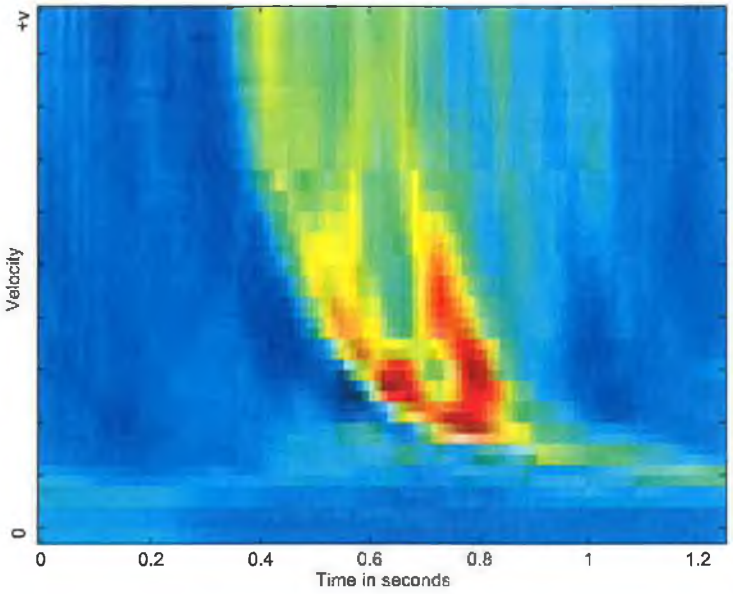


Figure 7.15: Hough transform of the cross-correlation array in Figure 7.14

believed that this is due to the vehicle body screening effects described in Section 3.3.2 creating a brief acoustical baffle between the tyre/road noise and the microphone array as it passes. Before and after the vehicle is adjacent to the microphone array, the vehicular sound is unhindered as it propagates to the microphone array. For the short time when the vehicle is close to the microphone, the vehicle itself is an obstruction to the noise that is primarily generated from underneath the body of the car and propagates at an upward angle towards the microphones. Without more precise acoustical measurements, the exact location of the distinct sound sources emanating from a passing vehicle is unknown. A vehicle with more than two axles displayed only two distinct patterns when in close proximity to the system.

Since it is often possible to visually distinguish front and rear vehicle noise in the cross-correlation array when in close proximity, it should be possible to detect these distinct sounds with an automated analysis of the data. Such information would contribute knowledge regarding the length, and therefore type of the vehicle. Recommended future work includes developing a pattern recognition technique that can not only detect a moving source, but also distinguish between axles. Naturally, individual axles from the same vehicle have the same velocity value and are a limited distance apart.

One of the difficulties with vehicle axle detection is distinguishing two separate vehicles in close proximity from a single long vehicle. Also, distinguishing axles first requires sharply defined and distinct evidence in the cross-correlation array in order to precisely define each axle and be confident of their relationship. It is currently difficult to ascertain the best-fitting model to apply to the cross-correlation array and would be a further challenge to match a pair of related models to a single vehicle. It must be first clarified what exactly the two individual cross-correlation shapes represent, by means of controlled experiments in a quiet environment. This is recommended as a future research direction.

7.5 Conclusions

The implementation of three different pattern extraction approaches has been described in this chapter; a sound amplitude-based approach and two cross-correlation approaches. The amplitude-based approach is used as a minimum benchmark to test the second two approaches, since it will never be robust enough to detect vehicles in noisy conditions. Both cross-correlation based methods have the common purpose of detecting and evaluating moving source behaviour based on evidence in a cross-correlation array.

The first cross-correlation method extracts protruding cross-correlation sequence peaks, and tracks their behaviour over time to investigate whether they represent a passing vehicle. Since the peak tracking method is based on tracking salient peaks, there is a danger that less noticeable events are overlooked and any early error in the process propagates to the end.

The second approach searches for regions or shapes of high correlation in the cross-correlation array that match the time-delay shape model of a passing vehicle. Most cross-correlation values within the shape region contribute to a decision regarding the presence of the shape in the cross-correlation array for certain parameter values, in a similar manner to Hough shape detection. Two different shape models are used; a rectangle and an S-shape matching the moving source model derived in Chapter 6. The latter shape was found to more accurately reflect the behaviour of a moving vehicle, and is therefore used during evaluation in Section 8.4.2.

There are merits to both cross-correlation approaches that can be evaluated based on accuracy of results and computational complexity. Chapter 8 describes the recording equipment, locations and reference data reliability. It then compares the different methods in terms of performance, accuracy and speed.

Chapter 8

Experiments and Results

8.1 Introduction

The automatic traffic monitoring systems developed in this thesis are evaluated in this chapter, based on a number of experiments using real traffic data. The equipment used to record audiovisual traffic signals is described, as well as the locations where the data was gathered. Experiments were performed to quantify the accuracy of the reference audio-based data against video evidence. The different methods are compared in terms of performance, accuracy and speed. Finally, an evaluation is performed of the methods based on experimental results.

8.2 Traffic recordings

In order to evaluate an audio-based traffic monitoring system, audiovisual traffic data was recorded on public roads and processed on a PC in the laboratory. It was not possible to find a suitable location to install a permanent recording system in close proximity to a road. Due to safety issues and power constraints, a recording system could not be left unattended at the roadside. This restricted the amount of traffic data gathered. During the course of 2 years, traffic data was recorded at 8 different locations. Audio and video signals of moving traffic were recorded at a range of locations with differing road types and background noise. These recordings were used as source data for traffic monitoring experiments, where the same algorithm parameters and thresholds were used for all experiments. The recorded audio signals were stored as standard PCM WAV-format files with 16-bit precision and a sampling frequency of 44.1kHz. The video files were recorded as AVI files and later converted to MPEG files for storage efficiency. Relevant measurements were made at each location, such as the distance between microphones and the distance to the road.

8.2.1 Recording equipment

The same recording equipment was used to collect audio traffic recordings at each location, however in some cases the geometrical distributions of equipment differed between recording locations. Where relevant, the measurements are indicated in the results. The traffic recording equipment can be summarised as a set of microphones and supporting structure, audio workstation, laptop and video camcorder.

The microphones used are a series of phantom-powered[1] Behringer ECM8000 condenser microphones with an omnidirectional polar pattern and linear frequency response. The *MOTU 896HD* audio workstation was used to provide phantom power for the microphones, digitise the audio signals and transfer the signals to a laptop to be recorded via a *FireWire* connection [2]. It is a 24-bit digital audio workstation capable of processing up to 8 microphone signals at a range of sampling frequencies up to 96kHz. A software program *n-Track Studio* was used to handle the multiple audio streams and write each channel simultaneously to individual files for storage. For some recording locations a *Sony DCR-PC100E* digital video camera was aligned behind one of the microphones to provide audiovisual evidence of recorded traffic. In a limited number of recordings, a volunteer used a hand-held GPS to measure the velocity of his car when passing the recording equipment. The GPS used was a *Navman 3000 GPS* sleeve on a Compaq iPAQ with Navman trip software. A reasonable estimate of the velocity accuracy of the GPS is ± 3 kilometers per hour.

The equipment used to record traffic data is of a much higher specification than might be expected in a low cost mass-produced system. The reason for this is so that the capabilities of audio analysis in the ideal case may be determined. It is another engineering task to investigate the lowering of accuracy if lower quality equipment is used.

[1] Phantom power is a DC voltage (11 - 48 volts) which powers the preamplifier of a condenser microphone

[2] FireWire (also known as i.Link or IEEE 1394) is a serial bus interface standard, offering high-speed communications and isochronous real-time data services.

8.2.2 Recording locations

Traffic data was recorded at different times at a range of locations. Over 3 hours of data from 5 different locations were used for testing, consisting of approximately 2,267 vehicles. These recordings included heavy traffic during rush hour and free-flowing traffic with regular intervals of silence. Furthermore, the recordings were performed in a variety of wind conditions and in proximity to potentially interfering sound sources. A variety of road types were recorded, from narrow lanes with low speed limits to dual carriageways. Table 8.1 lists the 14 traffic recordings used for experiments, during which audio and video signals of vehicular traffic were captured. To illustrate the amount and rate of traffic, the time duration, number of vehicles and average time between passing vehicle is noted for each file.

Table 8.1: Summary of traffic recordings used for experiments

File	Location	Duration (mins)	Quantity	Average time between vehicles
1	A	6.00	68	5.3s
2	A	13.02	126	6.2s
3	A	13.21	106	7.6s
4	A	23.59	155	9.3s
5	A	17.56	147	7.3s
6	A	20.18	149	8.2s
7	A	12.48	112	6.9s
8	B	7.28	49	9.1s
9	B	14.53	94	9.5s
10	B	16.55	118	8.6s
11	B	12.34	95	7.8s
12	C	21.28	463	2.8s
13	D	10.00	46	13s
14	E	20.24	539	1.5s
Total		3:31.6	2,267	

Type A recording

Files 1-7 recorded at location A listed in Table 8.1 were recorded adjacent to a 2-lane bidirectional public road beside an airport runway. Numerous airplane landings and take-offs were recorded together with vehicular traffic, a helicopter and an emergency

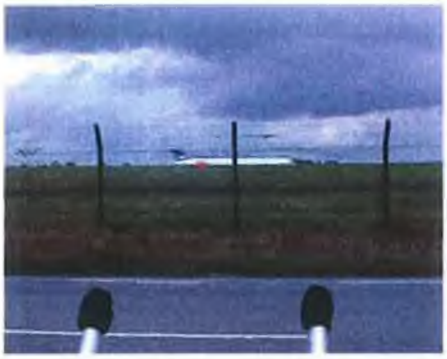


Figure 8.1: Image of type A recording location adjacent to the airport

vehicle with its siren activated. A video camera was used for files 1 and 2, placed perpendicular to the road and facing the airport runway. The view from the camera is visible in Figure 8.1. The average time between passing vehicles of all location A files was 7.24 seconds.

Type B recording

Type B files 8-11 were recorded adjacent to a 2-lane bidirectional public road near a train track in gusty wind conditions. The microphones were positioned in an identical fashion to recordings of type A. Passing diesel and electric trains were recorded together with vehicular traffic. No video signals were recorded. The average time between passing vehicle of all type B files was a vehicle every 8.7 seconds.

Type C recording

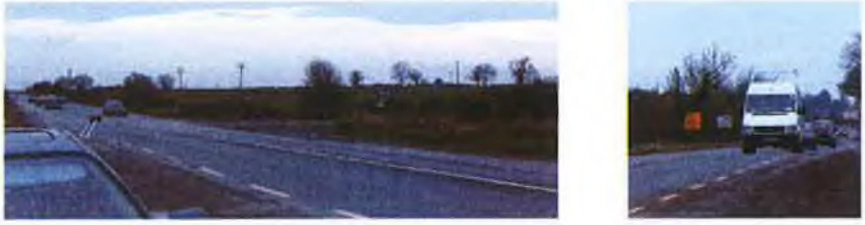


Figure 8.2: Images of type C recording location

File 12 of type C was obtained with the purpose of recording audio and video signals of heavy traffic. A 2-lane bidirectional public road was chosen with a speed limit of 110km/h. The microphones were positioned identically to type A recordings. Two video cameras were used facing both directions to record traffic approaching and departing the microphone array, as shown in Figure 8.2.2. High-density traffic was recorded at this location, where the average time between passing vehicles was a vehicle every 2.8 seconds. This is significantly higher than type A or B traffic. A single visual ground truth was obtained from comparing both video recordings and adjusting the time stamp to compensate for the difference between visual capture and projected time when located at the microphone array.

Type D recording

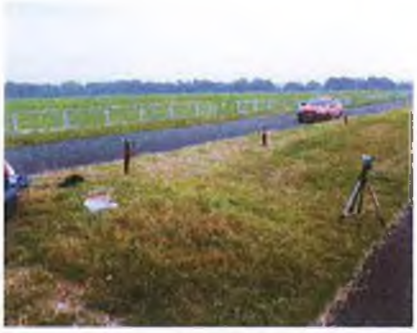


Figure 8.3: Image of type D recording location on a quiet road in a park

This location was chosen to record audio signals in a quiet setting from 3 widely spaced microphones, with no background noise and little wind. The outer left and central microphones were spaced 4m apart, the central and outer right microphone were spaced 5.5m apart. The data recorded is contained in file 13. A video camera was placed behind one of the microphones to record the visual information. Figure 8.3 illustrates the recording equipment on-site. The recording location was a quiet road in a public park with bidirectional single-lane traffic consisting of cars, bicycles, motorcycles and vans. Speeds varied between 30 and 80 kilometers per hour. There was a small amount of wind noise. The average time between passing vehicles was a vehicle every 13 seconds, which is relatively light traffic when compared to the other files.

A car was driven past the recording equipment multiple times at different known velocities, the velocities being measured by GPS and held constant during pass-by. In this manner, a limited number of recordings of a vehicle passing at a known velocity were recorded. This velocity data is described in Section 8.3.4.

Type E traffic recording

File 14 of type E was recorded adjacent to a wide 4-lane bidirectional public dual carriageway road with speed limits of 110km/h. The equipment and road is pictured in Figure 8.2.2. A single video camera was placed perpendicular to the road. Audio signals were captured by 6 microphones placed in a grid. The purpose of using 6 microphones was to obtain data appropriate for testing arrays of vertical and horizontally distributed microphones. The uppermost row of 3 microphones were elevated 180cm above the ground, each microphone in the grid being a distance of 20cm from the nearest vertical or horizontal adjacent microphone. The array was placed at a distance of 775cm from the yellow line at the side of the nearest traffic lane to the front of the microphones.

This recording location is particularly challenging, as it is a dual carriageway; 4 lanes in total, with 2 pairs of lanes in each direction separated by by a median strip of grass and partially covered by a low hedge. For clarity, the lanes are described as lane 1 - 4, with lane 1 being the lane farthest from the microphone array and 4 being the closest lane. Therefore, lane 3 and 4 contain traffic travelling in the same direction from right to left, and lane 1 and 2 contain traffic travelling in the same direction from left to right. The average time between passing vehicles was 1.5 seconds when considering all 4 lanes, or every 3.9 seconds for only the nearest 2 lanes. Similar to type C, the recording consists of high density traffic.

It was too dangerous to measure the road width due to the volume of traffic, therefore the physical road width was estimated. The distance between the microphone array

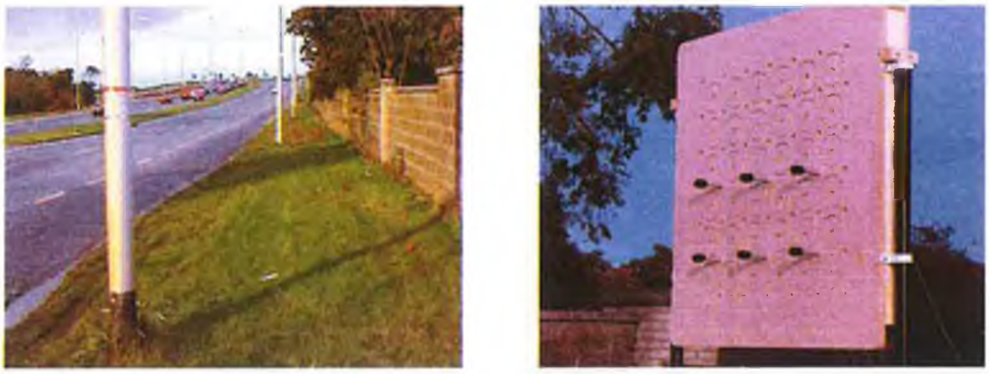


Figure 8.4: Image of type E recording location at a dual carriageway using a microphone array

and the farthest two traffic lanes was estimated at 10 - 17m based on a lane width of 3.5m and a median strip of width 2m [15]. The sound attenuation and time-delay resolution for a distance of 10m are such that it was impossible to monitor the farthest two lanes to any reliable degree of accuracy. It was decided to use the reference data based only on the nearest two lanes, as a fair test of the abilities of the traffic monitoring systems. Therefore, the type E reference data is henceforth only applicable to the nearest two traffic lanes in the dual carriageway.

8.3 Reference data based on audiovisual traffic recordings

Reference data (or "ground truth") listing passing vehicles was manually generated by the author for all recorded audio and video files separately. Each type of reference data was carefully created without reference to the other types of data. For example, the audio files were marked blind with no video or cross-correlation data available. The manual generation of data was unavoidably subjective, since it relies on a human auditory and visual perception of a vehicle in possibly noisy data. Every effort was made to be as objective as possible while generating reference lists of vehicles for each file. One of the measures taken was to avoid cross-checking different file types prior to completion of the reference data.

Cross-correlation arrays obtained from the audio data were visually observed to generate a third set of reference data. This is to manually determine evidence of known vehicles in the cross-correlation array. There are two reasons to obtain cross-correlation reference data. Firstly, the merit of the cross-correlation array in the task of detecting vehicles may be determined in this manner. Secondly, the accuracy of pattern extraction algorithms can be tested in automatically determining vehicle presence in the cross-correlation array.

When available, three lists were obtained for each recording session: audio, video and manual cross-correlation reference data. For the video and cross-correlation array, the vehicle direction was also noted. The different reference data types were time-synchronized using an abrupt clap that was performed at the start of each recording, in full view of the video recorder and in close proximity to the microphones.

The Venn diagram in Figure 8.5 illustrates the logical relationship used to compare

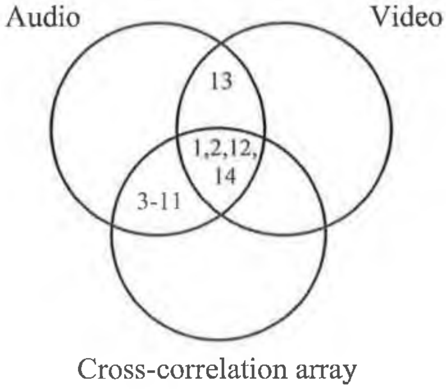


Figure 8.5: Venn diagram illustrating reference data overlaps

the different types of data. Ideally, passing vehicles would be detected in all data formats, thereby being represented at the centre of the diagram. Each reference data type was compared with the other data types to determine the overlap and reasons for discrepancies. Vehicles were not always detected by each reference data format, resulting in *missed events.* Imaginary vehicles may also be "detected", that are described as *false positives.*

8.3.1 Evaulation measures

Evaluations were carried out using precision and recall measures [187, 14]. Precision and recall are performance measures used to evaluate data returned by information retrieval systems [120, 161]. Precision-recall curves have been cited as an alternative to ROC curves [95]. ROC curves can present an overly optimistic view of an algorithm's performance if there is a large skew in the class distribution [46]. For this reason, it was decided to use precision-recall curves as evaluation measures.

For any given retrieved set of data, *recall* is the number of retrieved relevant (i.e. correct) items as a proportion of all relevant items. Recall is therefore a measure of effectiveness in evaluating performance, and can also be viewed as a measure of effectiveness in including relevant items in the retrieved set. For any given retrieved set, *precision* is the number of retrieved relevant items as a proportion of the number of retrieved items. Precision is therefore a measure of effectiveness in excluding non-relevant items from the retrieved set. The harmonic mean combines precision and recall into a single parameter for optimization. This is also known as the *F-measure*, or F_1 measure when recall and precision are evenly weighted. Harmonic mean tends strongly toward the least element of the list [43]. When compared to arithmetic mean it mitigates the impact of large outliers and aggravates the impact of smaller ones. In this manner, the harmonic mean is closer to the lower value than the arithmetic mean.

$$\begin{aligned} R &= \frac{\text{retrieved relevant}}{relevant}, && (8.1)\\ P &= \frac{\text{retrieved relevant}}{retrieved}, && (8.2)\\ F &= \frac{2}{\frac{1}{R}+\frac{1}{P}} = \frac{2PR}{P+R}, && (8.3) \end{aligned}$$

where P is precision, R is recall and F is the harmonic mean or F-measure of precision and recall. Optimal results would be to achieve 100% on both precision and recall, or F at the same time. However, the fundamental relationship between precision and recall necessitates a tradeoff when attempting to optimize both values. The F-measure can be used to summarize the effects of both precision and recall and is used when describing results. It falls in the range from 0 to 1, with 1 being the best possible score.

8.3.2 Comparison between audio and cross-correlation ground truth data

The audio and cross-correlation array data are compared in Table 8.2. For each file, the total number of events are given for each relevant region of the aforementioned Venn diagram in Figure 8.5. Events in region AC are correct results in the total subset. The precision, recall and F-measure for each file is given. This is to determine the merits of the cross-correlation array as a means for vehicle-detection.

The total F-measure in Table 8.2 and future results is not the average F-measure over all files, where each file is weighted equally. Rather, it is the total F-measure for all events across all files, as if every file has been concatenated into a single recording. For all recorded data, a total F-measure of 0.955 was obtained. This indicates that a very high proportion of vehicles are detected by both audio and cross-correlation data formats. This is to be expected, since both data formats are based on auditory information. Only 6 or 0.29% of 2042 audio events aurally detected are indistinguishable by the cross-correlation array.

Table 8.2: Comparison of audio and cross-correlation ground truths

File	AC	A	C	Precision	Recall	F
1	59	59	68	0.87	1	0.929
2	110	110	126	0.87	1	0.9322
3	98	99	106	0.92	0.99	0.956
4	144	146	155	0.93	0.99	0.957
5	137	137	147	0.93	1	0.965
6	140	140	149	0.94	1	0.969
7	108	108	112	0.96	1	0.982
8	45	45	49	0.92	1	0.957
9	93	94	94	0.99	0.99	0.989
10	110	111	118	0.93	0.99	0.961
11	87	88	95	0.91	1	0.951
12	432	432	463	0.93	1	0.966
14	473	473	539	0.88	1	0.935
Total	2,036	2,042	2,221	0.917	0.997	0.955

The number of cross-correlation detected events not aurally detected in the audio signal was 119 out of a total of 1450 audio events (8.2%) for all files except file 14. Including file 14 increases the number to 185 out of 2,042 (9.1%). In experiments based on location D data, it was found that the audio signals captured by microphones were largely uncorrelated and produced very poor results in source tracking. This was probably due to the large separation between the microphones. Therefore Type D data was not included in establishing the cross-correlation ground truth.

Upon analysis, reasons for the events present in the cross-correlation array not being aurally detected are as follows:

- closely sequential vehicles are “visible” in the cross-correlation data but not distinguished in the human-detected audio ground truth;
- simultaneous vehicles in different or identical directions are aurally indistinguishable;
- predominant noise from planes, trains or heavy goods vehicles mask quieter vehicles, resulting in them not being perceived by the human auditory system. However, sufficient traces of these vehicles are visible in the cross-correlation array to be detected.

Assuming all vehicles in the cross-correlation ground truth are correct detections, the manually generated cross-correlation ground truth is actually more accurate than the ground truth generated from human aural detection of passing vehicles. This further strengthens the argument for transforming the audio signals to create a cross-correlation array, since equivalent events are more accurately detected. Comparing the cross-correlation and audio data to the video ground truth indicates which data is a more accurate representative of passing vehicles.

8.3.3 Video ground truth compared to audio and cross-correlation data

The video ground truth is compared in Table 8.3 with the cross-correlation data, and in Table 8.4 with the audio data. The volume of data compared is 70.75 minutes in duration, with up to 1025 vehicles according to the video ground truth data.

In comparing results, file 14 is first excluded. For the reference video files 1,2 and 12, the cross-correlation F-measure is 0.929 whereas the audio F-measure is 0.842. The audio data (i.e. the manually generated ground truth based on human aural detection) is consistently less accurate than the cross-correlation data for each file and for a combination of the files, when compared to the video ground truth. File 14 is then included to increase the reference video files to 1,2,12 and 14. The cross-correlation F-measure is 0.73, while the audio F-measure is also 0.73. This provides evidence that the cross-correlation data becomes less reliable for location E data, obtained at a dual carriageway with a total of 4 lanes. For file 14 alone, the cross-correlation F-measure is 0.536 while the audio F-measure is 0.471. The cross-correlation data is

Table 8.3: Comparison of video and cross-correlation ground truths

File	Cross-correlation + video	Cross-correlation	Video	Precision	Recall	F
1	68	68	70	1	0.971	0.985
2	122	126	124	0.968	0.984	0.976
12	375	423	422	0.887	0.889	0.887
14	122	124	331	0.984	0.369	0.536
1,2,12	565	617	616	0.843	0.917	0.929
Total	687	935	947	0.735	0.725	0.730

Table 8.4: Comparison of video and audio ground truths

File	Audio + video	Audio	Video	Precision	Recall	F
1	69	69	79	1	0.873	0.932
2	105	110	126	0.955	0.833	0.889
12	333	398	422	0.837	0.789	0.812
13	46	46	46	1	1	1
14	103	106	331	0.972	0.311	0.471
1,2,12	507	577	627	0.879	0.809	0.842
1,2,12,13	553	623	673	0.888	0.822	0.854
1,2,12,14	610	683	979	0.893	0.623	0.73
Total	656	729	1025	0.899	0.64	0.748

still more accurate than the audio data, notwithstanding the drop in reliability for multi-lane dual carriageways or motorways.

It can be ascertained that the cross-correlation data is more closely aligned to the video data than the audio data. This provides evidence that the cross-correlation ground truth is actually a more accurate representation of passing vehicles than audio ground truth for these recordings. For this reason, the cross-correlation ground truth is used from this point forward as the reference data when testing audio-based traffic monitoring methods. 8.3% of passing vehicles observed in the video are not detected in the cross-correlation array when file 14 is excluded. By including file 14, this increases to 27.45%, reflecting the challenging environment of the dual carriageway recording in file 14.

8.3.4 Vehicle velocity ground truth data

In order to evaluate the ability of the audio traffic monitoring system to measure vehicle velocity, it is necessary to obtain a set of test cases where the vehicle velocity is known. However, using speed detection equipment at a public road without the correct procedure may provoke drivers to quickly alter their driving behaviour and increase the risk of accidents. The co-operation of the police and local traffic authorities is advised to use speed detection equipment on public roads. Such a large-scale operation in conjunction with local authorities was not possible for this research. Therefore an alternative solution was necessary to obtain known velocity measurements of passing vehicles. The method used was based on two stages; using a

GPS to determine the velocity of a moving vehicle, and using video data to calculate vehicle velocity.

A volunteer drove his vehicle past the microphones and video recorder at a constant velocity a number of times during the location D recording of traffic data. At the same time, a passenger in the car used a hand-held GPS to measure vehicle velocity, noting the speed of the vehicle at the time of passing the recording equipment. Public safety was a priority at all times, with all equipment being placed a measured distance from the side of the road. This procedure was repeated 6 times for different velocities. Simultaneously, the camcorder captured a video of the test car with a frame rate of 29.97fps. Since the length of the test car is known, the vehicle velocity can be determined by dividing the vehicle length by the time it takes the vehicle to pass a reference point. The number of frames required for the vehicle to pass the visual reference point in the scene are counted. Dividing by the frame rate returns the amount of time it takes for the vehicle to pass. In this manner, the velocity of the test car can be estimated from the recorded video.

Table 8.5: Velocity measurements of a known test vehicle in km/h, based on a hand-held GPS and video evidence

	GPS velocity	Video velocity	Difference km/h	% Difference
	41.86	43.65	1.79	4.1
	46.69	45.02	-1.67	-3.58
	61.18	63.54	2.36	3.71
	69.23	71.48	2.25	3.15
	66.01	63.54	-2.47	-3.89
	78.89	74.48	-4.41	-5.92
Average	60.64	60.28	2.44	4.05

To determine the accuracy of the video-based velocity estimation method, the test car velocities measured by the GPS and video are compared. The six cases of known velocity are listed in Table 8.5. The average difference between video and GPS velocity is approximately 2.44 km/h or 4.05%. For the higher velocity of 74.48 km/h, the error increases to a difference of 4.41 km/h or 5.92%. The video-based velocity estimation is therefore within a maximum of 4.4 km/h or 5.92% of the GPS velocity measurement for velocities below 79 km/h. Vehicle velocity can therefore be estimated from video recordings with a known frame rate and known vehicle length.

To build a larger video-based ground truth, a total of 150 test vehicles were chosen from Type A, C and E traffic recordings. These recordings were selected to represent a diverse range of recording situations; a quiet tertiary road, busy primary road and busy dual carriageway. The length of each test vehicle was individually determined from the manufacturer specifications of each vehicle model. The number of frames taken for each vehicle to pass two different visual reference point were counted and the average taken. Using the number of frames, vehicle length and video frame rate, the velocity was calculated for each of the 150 test vehicles to build a velocity ground truth with a calculated tolerance. This video-based velocity ground truth was used to compare against audio-based vehicle velocity measurements.

8.4 Automatic vehicle detection experiments

The ground truth data developed in Section 8.3 provides a reference of the quantity and characteristics of 2,267 individual vehicles over 3 hours of data from 5 different types of location. This reference data forms the basis of experiments, testing and comparing the three automatic traffic monitoring techniques described in this thesis. Each system is individually evaluated in Section 8.4.1 to 8.4.3 and their performances compared in Section 8.5.

8.4.1 Vehicle detection using acoustic amplitude

The sound amplitude-based vehicle detection method described in Section 7.1 was tested to determine its accuracy. Table 8.6 lists the overall accuracy in detection for each file. 14 recordings containing a total of 2,267 vehicles of 3:31.7 duration were used. The cross-correlation ground truth was selected as the reference data since it is a more accurate representation of the video data and includes more events than the audio ground truth. Furthermore, the video ground truth is not based on acoustic information and only a small subset of the traffic recordings include video data.

When two or more vehicles are simultaneously present it is impossible to distinguish individual vehicles based on measured acoustic amplitude alone, resulting in a number of missed vehicles. As described in Section 7.1, the method includes an amplitude threshold based on a percentage of the average amplitude. If the threshold value is too high vehicles are missed, but if the threshold is too low there are an excessive

Table 8.6: Vehicle detection accuracy of the volume-based traffic monitoring system

File	Precision	Recall	F
1	0.417	0.435	0.426
2	0.524	0.516	0.519
3	0.433	0.576	0.494
4	0.287	0.489	0.362
5	0.357	0.428	0.389
6	0.418	0.5	0.455
7	0.417	0.486	0.449
8	0.725	0.725	0.725
9	0.591	0.553	0.571
10	0.661	0.574	0.614
11	0.374	0.419	0.395
12	0.628	0.609	0.619
13	0.40	0.739	0.519
14	0.512	0.793	0.622
Total	0.499	0.595	0.535

number of false detections. The results in Table 8.6 are based on the threshold value necessary to achieve the best F-measure for the captured data. The total precision and recall values over all the files were 0.49 and 0.56 respectively, with a total F-measure of 0.535. Since the location of a vehicle is not measured from the acoustic amplitude information, it is not possible to determine vehicle direction or velocity. Therefore velocity experiments were not performed using the sound amplitude-based vehicle tracking approach.

It is clear from the results that an acoustic amplitude-based approach is an unreliable method for traffic monitoring, detecting approximately only 50% of vehicles. A large number of vehicles were missed and there was a significant number of false detections. Nevertheless, the amplitude-based approach did succeed in detecting vehicles in the right conditions. These conditions include relatively little background noise, successive vehicles passing at a large distance from each other and generating a sufficient amount of noise to be distinguished.

A simple sound amplitude test could be used to act as a trigger to a series of more advanced localization-based signal processing approaches to track vehicles. When there is no traffic present, there is no need for sound analysis that requires significant processing or power, therefore the system could return to "sleep" mode while the sound amplitude indicator continues to monitor activities. Therefore information based on sound amplitude can play a role in traffic monitoring, although it is not sufficiently reliable to detect vehicles without further signal analysis.

8.4.2 Cross-correlation shape matching-based pattern extraction

The automatic shape matching cross-correlation pattern extraction method described in Chapter 7 is now examined for accuracy, when compared to the cross-correlation ground truth data. Two different shape models were developed - a rectangle and an S-shape matching the moving source model derived in Chapter 6. The latter shape described in Section 7.3.2 more accurately reflects the behaviour of a moving vehicle, and is therefore used during experiments. It is first evaluated based on its ability to detect passing vehicles. Following this, the accuracy in automatically measuring vehicle velocity is evaluated.

Vehicle detection accuracy

The model-based shape matching pattern extraction approach described in Section 7.3.2 was tested for accuracy in detecting vehicles over 13 different audio recordings. The results for each recording file are shown in Table 8.7. It was necessary to set 3 thresholds in the algorithm to automatically extract vehicles from the parameter space, as described in Section 7.3.2. The selected values of these thresholds directly influence the accuracy of vehicle detection. Depending on the preferred results, the thresholds can be chosen with a view to optimizing precision, recall or both. Typical systems seek to optimise both precision and recall, thereby maximising the F-measure. The precision and recall values for different threshold values applied to the type C recording are listed in Table B.1 and illustrated in Figure B.4 in Appendix B. The threshold values were selected to maximize the F-measure for each file, where the results for each file are shown in Table 8.7.

The overall F-measure was 0.927 for 2,196 vehicles over 13 recordings of 3 hours 31.7 minutes. This high overall F-measure provides strong evidence that the shape matching pattern extraction approach is a highly accurate approach to detecting vehicles. Only 8% of vehicles were missed and only 1.6% of detected vehicles were

Table 8.7: Shape matching pattern extraction results compared to the cross-correlation ground truth

File	Match	Ground truth	Shape matching	Precision	Recall	F
1	53	69	54	0.982	0.768	0.862
2	109	130	119	0.916	0.839	0.876
3	98	105	103	0.952	0.933	0.942
4	148	153	155	0.955	0.967	0.961
5	134	143	141	0.950	0.937	0.944
6	142	148	150	0.947	0.959	0.953
7	98	106	110	0.891	0.925	0.907
8	41	46	47	0.872	0.891	0.882
9	89	90	99	0.899	0.989	0.942
10	114	115	126	0.905	0.991	0.946
11	86	92	93	0.925	0.935	0.930
12	439	463	454	0.967	0.948	0.957
14	468	536	509	0.919	0.873	0.896
Total	2,019	2,196	2,160	0.935	0.92	0.927

false positives. Such highly accurate results are reflected across all the tested files, with the difference between lowest and highest F-measure being only 0.09 for the same thresholds being used for every file. The large quantity of 2,196 events in the reference data increases confidence that the shape matching pattern extraction approach is a highly accurate method for automatically detecting vehicles.

One early concern was how the system would handle simultaneous vehicles. As described in Chapter 6.2, the road length being observed was a maximum of 6m, or a length of 2m for a range of observation angles between -45 and 45 degrees, due to the parameters chosen. Since the average length of a typical car is 4.25m, it is unlikely that more than a single vehicle could occupy the observed road length in a single lane at the same time. Therefore, for a road with two lanes, there is a very small probability that two or more vehicles will pass simultaneously. As expected, there were very few recorded cases out of the 2,196 events when two or more vehicles passed simultaneously. There was normally a small difference in the time and speed of passage. The shapes of two simultaneous vehicles travelling in different directions were generally both detected in the parameter space. Examples are shown in the cross-correlation images displayed in Section B.1 of Appendix B. When two vehicles occurred simultaneously in the cross-correlation array, they were both represented in the parameter space and detected as separate events.

Another question was how the system would cope with loud noise that may possibly mask the road traffic. In the case of recordings from location A, the sounds of numerous airplanes landing and taking off were present in the audio signal. A moving source pattern was visible in the cross-correlation array due to many of these airplanes. As a result of the prolonged high amplitude of sound emitted by an airplane, the pattern from an airplane in the cross-correlation array stretches over a far greater time period than for a road vehicle, while changing gradually in value of τ. This is evident in cross-correlation images displayed in Section B.1 of Appendix B. For a very short time a small number of passing vehicles were acoustically masked and therefore not detected in the cross-correlation array. This was primarily during the loudest noise generated by the aircraft as it landed or took off, accounting for a number of the missed vehicles in files from location A. However, a similar amount of vehicles were detected in the cross-correlation array, despite the significant presence of aircraft noise in the background. An identifiable pattern in the cross-correlation array due to a vehicle could be distinguished from cross-correlation noise due to any aircraft.

Velocity and direction accuracy

The shape matching pattern extraction approach was tested for accuracy in estimating vehicle velocity. 150 test cases of passing vehicles were selected from 3 different recording locations (A, C and E). The vehicle velocities of the test cases were carefully measured using the video-based method described in Section 8.3.4. The measured velocities were found to range from 33 km/h to 76 km/h. These measured velocities were compared against automatically estimated vehicle velocities using the shape matching pattern extraction approach.

Figure 8.6 displays a scatter plot representing the relationship between measured and automatically estimated vehicle velocities for a range of test samples. The solid red line illustrates ideal results where zero error occurs, i.e. the measured value equals the estimated value. The further a data point is from the solid red line, the greater the velocity estimation error. The solid black line displays the least-squares line of best fit of the actual data samples. It can be observed from Figure 8.6 that there is a general correspondence between measured and estimated vehicle velocity; the

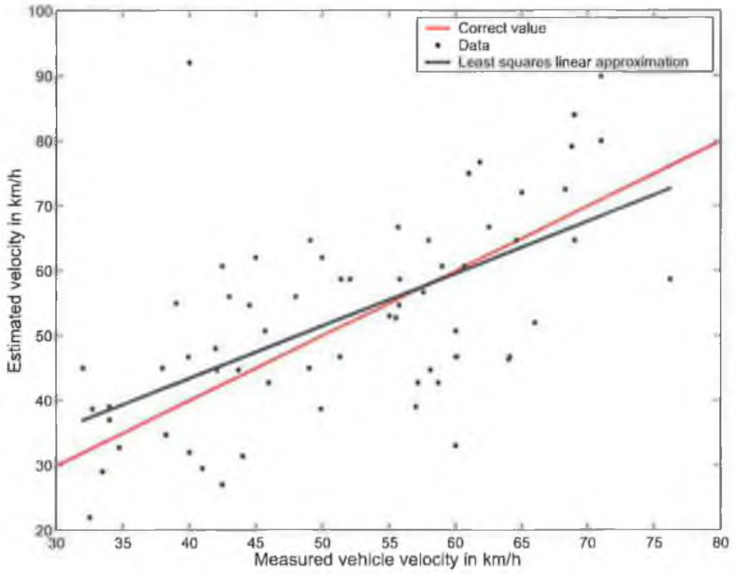


Figure 8.6: Comparison between measured and automatically estimated vehicle velocity

higher the vehicle velocity, the higher the velocity estimate. The estimations both under and over-estimate the true vehicle velocity for a range of error values. The magnitude of these error values is displayed in Figure 8.7 in km/h over the range of measured velocities. Figure 8.8 displays the velocity error as a percentage of the true velocity. For the test cases examined, the maximum velocity error value is measured as 19km/h, while the maximum percentage velocity error is 42.8%.

In Section 6.3.3, the maximum theoretical accuracy of vehicle velocity for the set of parameters used by the audio system was described as ±5 km/h. This is based on a sampling frequency of 44.1kHz and interpolation factor of 4. It is therefore expected that the velocity measurement will be less than or equal to the theoretical accuracy of ±5 km/h. From Section 8.3.4, the error in reference velocity measurements is estimated at ±2km/h, increasing the overall theoretical velocity error to ±7 km/h. The actual velocity accuracy was measured at ±19km/h, a significantly larger error than the optimal theoretical error of ±7km/h. A number of reasons for the difference between actual and theoretical velocity accuracy are now described.

The theoretical velocity accuracy does not take into account the presence of noise in the audio data, false peaks in the cross-correlation sequence or errors due to the pattern analysis technique. Consider Figure 7.14, where a model is superimposed on the typical signature of a passing vehicle in a cross-correlation array. Since the

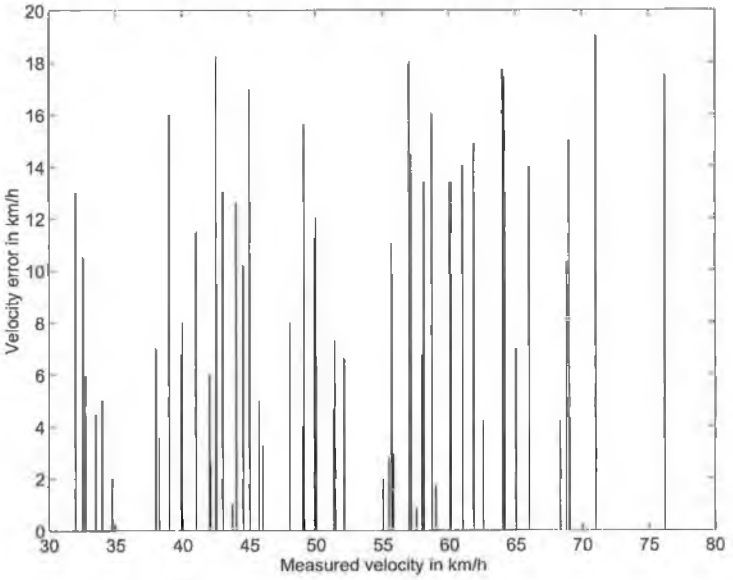


Figure 8.7: Accuracy of shape matching-based velocity estimation in km/h

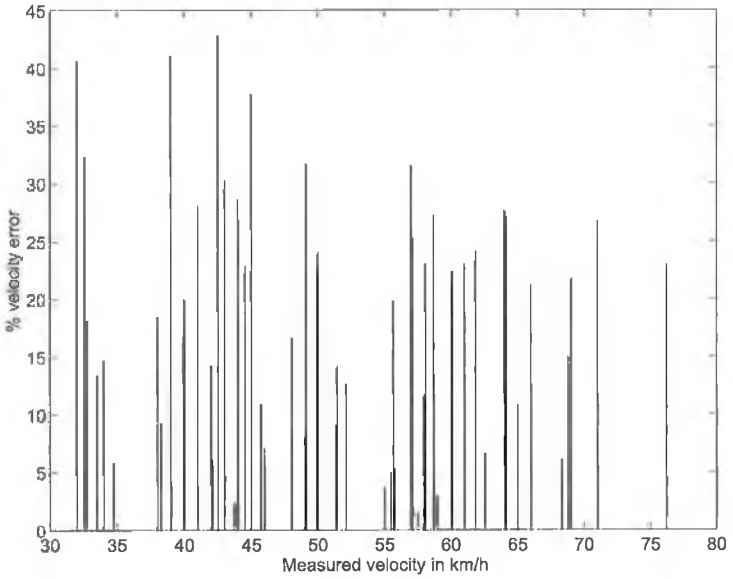


Figure 8.8: Accuracy of shape matching-based velocity estimation as a percentage of measured velocity

array data is broader and less defined than the model, it is possible to superimpose multiple models with slightly different velocity values and reference times. This results in a localized high-value region in the parameter space, as illustrated in Figure 7.15. More than one unique velocity and reference time fits the data with high probability, meaning that a number of different results are equally valid according to the parameter space and cross-correlation array. Based on the cross-correlation data being used, it is practically impossible to define the most appropriate model, since a number of different models may be suitable. Therefore, the accuracy in determining the source velocity is constrained by the sharpness of the cross-correlation array.

If there are multiple vehicles passing at the same time or correlated noise is present, the signature of a moving source in the cross-correlation array becomes interspersed with false peaks and competing correlation values. The pattern analysis approach is designed to tolerate such noise and lack of information, maintaining a high level of vehicle detection despite such problems. However, such noisy and confusing correlation data increases the difficulty of optimally fitting the correct model. In such cases, a similar model may fit better than the true model, resulting in a higher velocity error.

In summary, the accuracy achieved in measuring vehicle velocity for a limited number of test cases of vehicles travelling between 33 and 76 km/h is up to 19 km/h or 42.8%. Based on calculations in Section 6.3.3, it is expected that this error would increase for higher velocities. One can conclude that the velocity accuracy of the shape matching cross-correlation traffic monitoring system is sufficient to provide an indication of the general speed of vehicles, but is not accurate enough for precise speed measurements.

8.4.3 Peak tracking based cross-correlation pattern extraction

In this section, experimental results from the automatic peak-tracking based pattern extraction method described in 7.2 are examined for accuracy with respect to the cross-correlation ground truth data. The peak-tracking method is first evaluated based on its ability to detect passing vehicles. Following this, vehicle velocity accuracy is measured.

Vehicle detection accuracy

Automatically generated results using the peak-tracking method are compared with the cross-correlation ground truth in Table 8.8. The overall F-measure is 0.737, while the total precision and recall values are 0.692 and 0.789 respectively. These values were obtained for a total of 2,196 vehicles over 13 recordings of 3 hours 17.19 minutes duration.

Table 8.8: Peak tracking-based pattern extraction results compared to the cross-correlation ground truth

File	peak and GT	peak	GT	Precision	Recall	F
1	64	95	69	0.674	0.928	0.781
2	119	177	130	0.673	0.915	0.776
3	86	102	105	0.843	0.819	0.830
4	113	142	153	0.796	0.739	0.766
5	104	166	143	0.627	0.727	0.673
6	108	141	148	0.766	0.73	0.747
7	72	118	106	0.610	0.679	0.643
8	46	75	46	0.640	1	0.703
9	89	176	90	0.506	0.988	0.669
10	114	220	115	0.518	0.991	0.680
11	91	178	92	0.511	0.989	0.674
12	310	390	463	0.795	0.669	0.72
14	417	524	536	0.796	0.778	0.787
Total	1733	2504	2196	0.692	0.789	0.737

It can be observed that the F-measure ranges from 0.643 to 0.830 with a total F-measure of 0.737. There is a reasonable accuracy in vehicle detection, when compared with the acoustic amplitude-based method. However, the accuracy is significantly lower than the shape-matching cross-correlation approach. The cases where vehicles were not detected can be explained by a number of different reasons, listed as follows:

1. Relevant cross-correlation peaks were occasionally missed, increasing the difficulty in starting or continuing to track a pattern over successive cross-correlation sequences;
2. Cross-correlation peaks were occasionally linked to an incorrect trail, causing the trail not to be a true representation of the passing vehicle it supposedly

represents;

3. A cross-correlation peak is linked to an otherwise correctly shaped trail, redirecting the trail in the wrong direction and causing the trail not to be a true representation of the passing vehicle it supposedly represents;
4. Errors in matching are propagated through time to result in an inaccurate trail;
5. single vehicles are erroneously detected as two individual trails and mis-classified as two separate vehicles, increasing the number of false positives;
6. A trail is not long enough to be able to consider it as a passing vehicle, even though it is early evidence to that effect;
7. The trail is not long enough to be able to match a model with reasonable accuracy.

The reason for exploring the peak-tracking approach was to avoid the significant overhead of memory storing a large cross-correlation array in the context of a very low cost system requirement. By extracting relevant peaks and tracking their behaviour over time, the rest of the cross-correlation sequence may be immediately discarded. However, results and failure cases demonstrate that this approach is flawed. There are too many steps where small input disturbances can radically alter the outcome, from the correct detection of all cross-correlation peaks and the linking of peaks to correct trails, to the correct matching of a model to a completed trail. From the moment that an error is introduced it propagates through the data to result in a false or missed detection or inaccurate measurement of the time of passage.

The peak-tracking approach detects, link and tracks cross-correlation peaks over time in order to match a moving source model and determine the parameters of a passing vehicle. It has presented an F-measure of 0.737 based on 2,196 vehicles. It falls to the user to decide whether this detection accuracy is tolerable.

Velocity accuracy

The peak tracking-based pattern extraction approach was tested for accuracy in estimating vehicle velocity. 150 test cases of passing vehicles were selected from 3 different recording locations (A, C and E). The vehicle velocities of the test cases were carefully measured using the video-based method described in Section 8.3.4.

The measured velocities were between 33 and 76 km/h, with an average of 55.4 km/h and standard deviation of 11.57. These measured velocities were compared against automatically estimated vehicle velocities using the peak-tracking approach.

Figure 8.9 displays a scatter plot representing the relationship between measured and automatically estimated vehicle velocities for a range of test samples. The solid red line illustrates ideal results where zero error occurs. The further a data point is from the solid red line, the greater the velocity estimation error. The solid black line displays the least-squares line of best fit of the actual data samples. It can be observed that the best-fit line for data samples is almost horizontal and not linearly increasing as it should be. The estimated vehicle velocity bears little relation to measured velocity in most cases. The estimations both under and over-estimate the true vehicle velocity for a range of error values. The magnitude of these error values is displayed in Figure 8.10 in km/h over the range of measured velocities. Figure 8.11 displays the velocity error as a percentage of the true velocity. For the test cases examined, the maximum velocity error value is measured as 209km/h, while the maximum percentage velocity error is 400%.

The peak-tracking approach to vehicle velocity measurement returns unreliable velocity results that often differ greatly from the true value. By its nature, the peak-tracking approach is prone to selecting an incorrect model, as was described in Section 8.4.3. This can be caused by a number of reasons; an insufficient range of points along the shape increases the number of models that match the data, thereby increasing the error. Incorrectly linked points may cause an estimated shape to follow an incorrect path, presenting a false indication of the true shape. An event may prematurely die due to transitional noise or absence of a distinct correlation peak, making it difficult to fit a model. Any combination of these problems increases the difficulty of selecting the best fitting model. From an analysis of the velocity results, one can conclude that the velocity accuracy of the peak-tracking system is insufficient to approximate the general speed of vehicles.

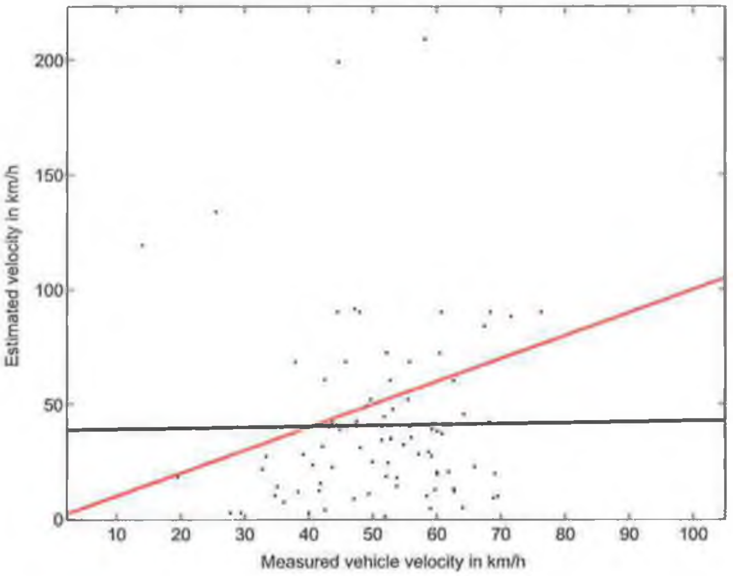


Figure 8.9: Accuracy of peak tracking-based velocity estimation in km/h

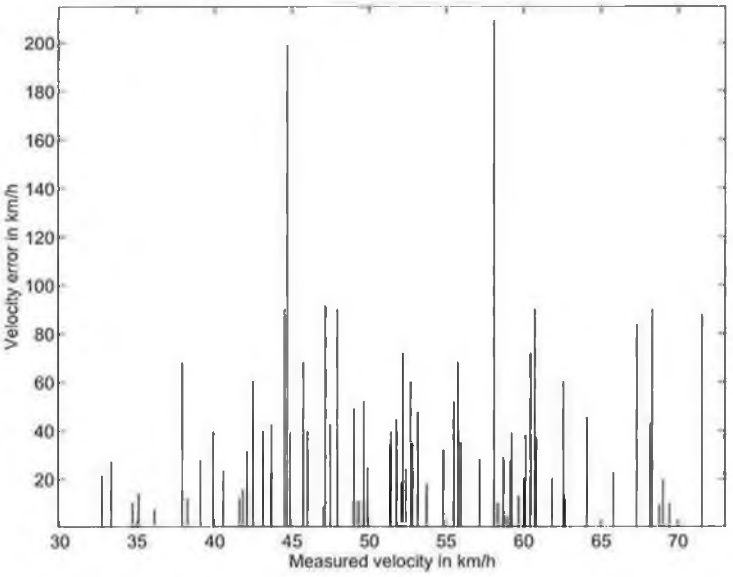


Figure 8.10: Accuracy of peak tracking-based velocity estimation in km/h

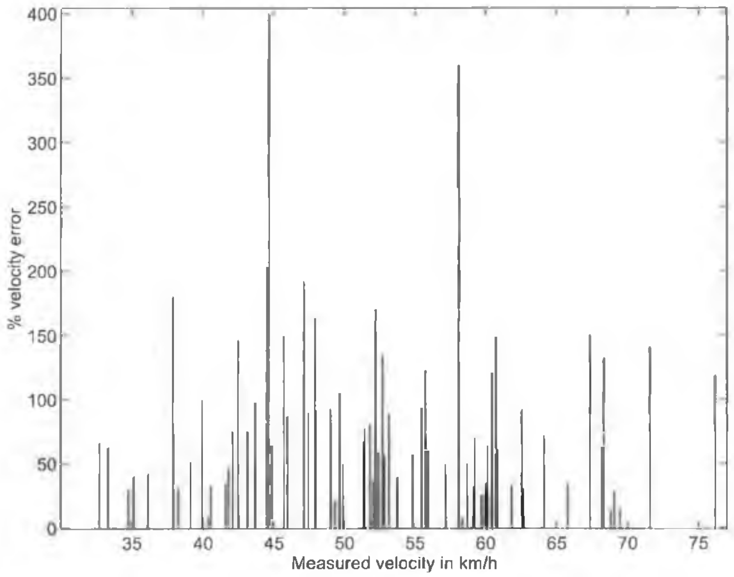


Figure 8.11: Accuracy of peak tracking-based velocity estimation as a percentage of measured velocity

8.5 A comparison of automatic traffic monitoring systems

8.5.1 Processing speed of traffic monitoring systems

The performance of a system is based not only on its accuracy, but also the time, resources and cost of performing the task. Therefore, the processing speed of each system is now compared. Monitoring traffic based on audio information cannot provide instant results, as it requires an analysis of the audio signals over a limited time period. By buffering a certain quantity of the audio signal, the traffic data could be processed in real-time and results presented after a fixed time set by the length of the buffer. The required buffer size depends on the values of O_w, f_s, m and the lowest possible velocity considered, since it determines the length of the slowest τ model. The computational requirements of the different systems described so far are examined here.

Experiments were performed using a Dell Precision 330 with Intel Pentium 4 Processor, CPU speed of 1.5GHz and 261MB of RAM. The algorithms were programmed for Matlab version 6 and run on the Windows 2000 operating system. A more precise timing calculation is desired, where the algorithms are optimised and the steps are quantified in terms of the number of multiplications, additions, iterations etc. In this manner the timing calculations would be more appropriate when compared, and is therefore a recommendation for future work. Table 8.9 compares the time taken to run each automatic traffic monitoring method based on 16-bit PCM audio wav files with a sampling frequency of 44.1kHz. The audio files used to evaluate processing speed consisted of 8,685 seconds of audio data containing 1,371 vehicles.

Table 8.9: Comparison of the computational time taken to analyse audio data in automatically detecting vehicular traffic

	Sound amplitude	Shape matching	Peak tracking
Total Computational time	155.59	15,879	69,375
Computational time to process 60s of data	1.075	109.699	479.294
Computational time to process 1s of data	0.018	1.828	7.988

The acoustic amplitude-based method is clearly the fastest approach, taking 1.8 seconds to process 1 minute of an audio signal, i.e. significantly faster than real-time. One of the reasons for the speed of the sound amplitude-based method is that it does not need to compute the cross-correlation data. The slowest method is the peak-tracking approach, which requires 7.9 seconds to process 1 second of audio data. The peak-tracking method is 4.4 times slower than the shape matching-based method, which is in turn 102 times slower than the amplitude-based method.

To examine the processing speed in further detail, Figure 8.12 displays the relationship between processing time for a single window of data and number of samples. The number of samples is determined by the window size and sampling frequency. The processing speed is subdivided into different algorithm stages (i) read audio data, (ii) window the data and (iii) obtain the cross-correlation sequence. Within a single window of the range of sizes analysed, the time taken to read audio data and windowing are independent of window size. However, the time taken to calculate the cross-correlation sequence is directly related to window size, showing a stepped result as opposed to the expected linear increase. Visible in the diagram is a stairs effect, or series of sudden jumps. This jump is due to the use of a fast-fourier transform and fixed window length in the process of obtaining the cross-correlation sequence.

The time taken to process an audio file depends on a number of factors. These

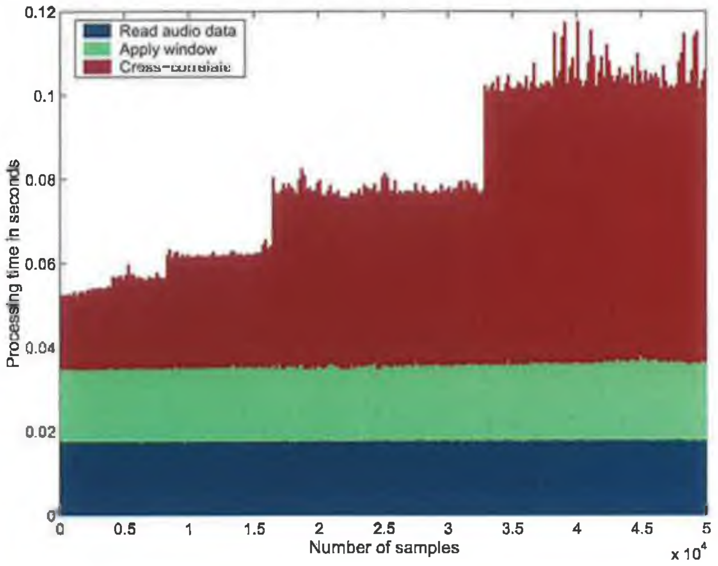


Figure 8.12: Processing time duration for increasing number of samples in audio data. The processing time is broken down into stages; read audio data, apply Hamming window and cross-correlate audio signals

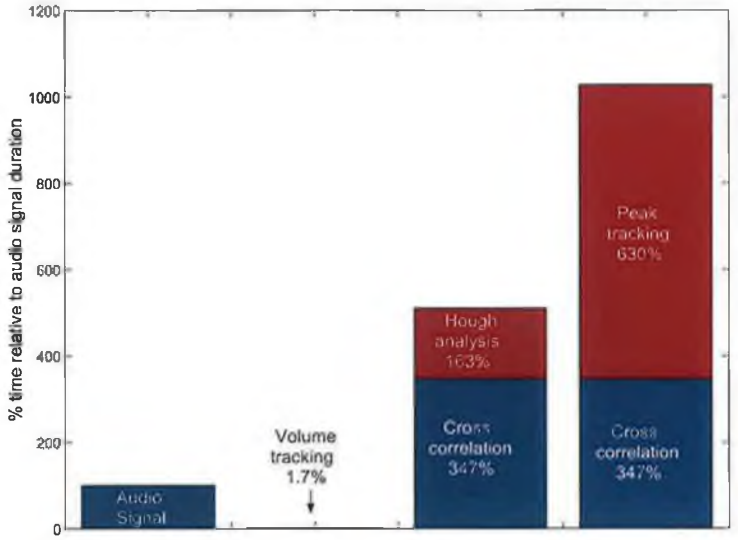


Figure 8.13: Processing time duration of different methods as a percentage of the audio signal time length, fs=44.1kHz, smallStep = 500, bigStep = 5000, width = 81, D = 5, m = 0.15

include the sampling frequency, window size, window overlap or hop size, length of audio file and the type of processing to be implemented. Assuming the audio signal is windowed and processed on a window-by-window basis, Equation 8.4 can be used to calculate the number of iterations i, based on these factors.

$$N_i = round\left[\frac{N_a - L_w}{O_w} - 1\right], \tag{8.4}$$

where N_i is the number of iterations, N_a is the number of audio samples, L_w is the window length and O_w is the window overlap length. A measurement is made to determine how long it takes to process a single window of data. Using this measurement, the time duration for a single window multiplied by the number of iterations N will result in the time required to process a known section of audio data.

8.5.2 Summary of system performances

In order to evaluate the different audio-based systems presented in this thesis, each system is compared in Table 8.10. The method based on sound amplitude is the fastest approach, however it results in the worst count accuracy. It has difficulty in distinguishing multiple vehicles that pass in close proximity and is overwhelmed by loud background noise.

Both the peak tracking and shape matching approaches use cross-correlation data to detect vehicles. Due to this, they are reasonably robust to background noise and are capable of distinguishing multiple vehicles. The peak tracking method has a better count accuracy than the sound amplitude approach and requires far less memory to archive cross-correlation data. However it is the slowest method and

Table 8.10: Performance comparison of automatic audio-based vehicle detection methods

	Sound amplitude	Shape matching	Peak tracking
Processing speed for 60 seconds	1.1	109.7	479.3
Overall Detection F-measure	0.51	0.93	0.75
% deviation from true count	0.51	8.06 of 2196	16.4 of 2089
Maximum velocity error (km/h)		19	209
Maximum % velocity error		42.8	400

presents unreliable velocity measurements. The shape matching-based method is slower than the approach based on acoustic amplitude and moderately faster than the peak tracking approach. Its primary advantage is that the shape matching-based method returns a highly accurate vehicle count. The velocity results are reliable, even if the accuracy tolerance is broader than desired. For these reasons, a shape matching-based pattern extraction algorithm applied to a cross-correlation method is the preferred method to monitor traffic.

8.5.3 Comparison between the shape matching traffic monitoring system and existing traffic monitoring technologies

In this section, a comparison is made between the audio-based shape matching system and existing traffic monitoring technologies described in Chapter 2. Reference data and experiments evaluating the systems are different, therefore any comparison must be treated with a degree of caution. Nevertheless, examining the performance of existing traffic monitoring technologies provides an indication of what performance the family of related technologies operate in.

Two evaluation results are used; the Minnesota Department of Transportation evaluation described in Section 2.2.2 and the Texas Transportation Institute sensor evaluation in Section 2.2.4. Both evaluations reported traffic sensors measuring vehicle count accuracy as a percentage of the correct number of vehicles. According to the Minnesota Department of Transportation evaluation, the video and passive acoustic devices were found to count with an error of between 4 and 10% of baseline traffic volume data. Pulse ultrasonic, doppler microwave, radar, passive magnetic, passive infrared and active infrared were found to count with an error of within 3% of the baseline. In the Texas Transportation Institute sensor evaluation, video system count error was within 10% until speeds dropped below 40mph, when the count error increased to 10 to 25%. The radar system count error was always within 10%. When speeds were over 40mph, the beamforming-based acoustic sensor count error was within 10% and with slow speeds it rose to 32%.

Based on a reported count accuracy of 10 - 25% for established audio and video sensors, the cross-correlation shape matching traffic monitoring system compares favourably with an overall count error of 8%. This is with a similar quantity of test data, since the Minnesota Department of Transportation evaluation is based on 1,923 vehicles.

The Minnesota Department of Transportation and the Texas Transportation Institute evaluations also measured accuracy in velocity measurements. According to the Minnesota Department of Transportation evaluation, all the devices were within 8% of the baseline speed data, with radar, doppler microwave and video being the most accurate. In the Texas Transportation Institute sensor evaluation, the video system speed estimation was with an error of between 0 and 5mph. The radar system speed accuracy was excellent except when speed dropped below 20mph. The velocity accuracy achieved in the author's work by the cross-correlation shape matching system of errors up to 42.8%, was a significantly worse result than existing technology accuracies of within 8%.

To the extent that different evaluations may be compared, it appears that the velocity accuracy of the cross-correlation shape matching method is lower than that of existing traffic sensors. This is primarily due to the high sensitivity of velocity estimation to the accurate measurement of the slope model. However, the cross-correlation shape matching method can achieve equivalent vehicle count accuracy as existing traffic sensors. This is with a far more economical system using two microphones as opposed to a large array of microphones or cameras that require precise calibration.

8.6 Conclusions

This chapter described a range of experiments evaluating each automatic traffic monitoring system developed during the course of this project using a large quantity of real traffic data in a variety of recording locations. The range of chosen recording environments were described, together with the type and quantity of data in the recording files. The recording equipment used to capture audiovisual traffic signals was identical for each recording location, although the geometrical parameters varied slightly.

When available, three sets of reference data were manually generated for each recording: aurally detected audio events as well as video and cross-correlation based on visual analysis. In this manner, the accuracy of the cross-correlation array was evaluated prior to evaluating the accuracy of pattern extraction algorithms. A very high proportion of vehicles were detected by both aurally detected events and cross-correlation data formats. The cross-correlation ground truth was found to be more accurate than the aurally detected ground truth and was used as the reference data for experiments. Over 3 hours of data containing 2,267 individual vehicles from 5 different types of location were used to test each traffic monitoring approach.The lengths of recognised vehicles and video data were used to build a vehicle velocity ground truth. This was used to test the accuracy of system velocity estimations.

The three audio traffic monitoring systems were evaluated: acoustic amplitude, tracking cross-correlation peaks and shape detection in the cross-correlation array. Each system was compared based on accuracy, speed and storage requirements.The approach based on sound amplitude was extremely fast and efficient, however it returned the lowest detection accuracy and was unable to estimate vehicle velocity. The peak-tracking method did not have large memory requirements and demonstrated a higher accuracy than the method using acoustic amplitude. Despite this, it displayed the slowest processing speed and highly inaccurate velocity estimations that could not be relied upon. Finally, the shape matching approach gave the most accurate vehicle detection result together with reasonably accurate velocity estimation. Although it requires storage of a section of cross-correlation array, it was significantly faster than the peak tracking method. When compared with existing traffic sensor technologies, it indicated the potential of returning equivalent detection accuracies. Therefore, the shape matching pattern extraction method applied to the cross-correlation array is the audio-based traffic monitoring method of choice.

Chapter 9

Conclusions

9.1 Summary and observations

This thesis has proposed the passive monitoring of vehicular traffic by means of acoustical data. The system developed by the author consists of up to two slightly separated microphones and signal processing capabilities, situated perpendicular to the road.

A description of existing traffic sensors and their evaluated performances was presented. There are a large range of traffic sensors available, based on a variety of technologies. No single sensor was reported to perform optimally in all conditions and according to all criteria, during a series of substantial comparative evaluations reported in the literature. Therefore, the optimal traffic sensor depends on the traffic monitoring environment and purpose for retrieving traffic data. Only two systems utilizing traffic-generated acoustical signals is currently commercially available. Based on the information available, it uses a computationally intensive beamforming approach and by necessity a large array of microphones. This thesis describes the development and evaluation of systems based on an efficient time-delay of arrival source localization technique using 2 microphones.

Described in Section 4.4, there are two audio-based products that are currently available for basic traffic monitoring using beamforming; SmartSonic by IRD Inc. and SAS-1 by SmarTek Systems. The technology behind the system is described in Section 4.4.1. Relevant research literature is described in Section 2.1.9. The estimation of vehicle speed and position using a single sensor was attempted by Couvreur and Bresler [45] using the Doppler effect. Pérez-Gonzáles and López-

Valcarce published a series of papers describing vehicle velocity estimation using the time delay between a pair of microphones [150, 113, 112, 111]. The use of wideband array processing algorithms for acoustic tracking and classification of ground vehicles, such as army tanks, is described by Pham et al [152, 151]. Forren and Jaarsma [62] describe a tyre-noise based traffic monitoring approach using a microphone array to localize the sound source by means of cross-correlation. An array-based traffic monitoring technique applied to urban situations was described by Chen et al. [38, 39] which also uses a cross-correlation based algorithm. Similar to Forren, Chen did not extract the traffic indicators automatically from the data but relied on manual intervention. Nevertheless, the cross-correlation approach described by Forren and later Chen is closely aligned to work described in this thesis.

In short, a limited number of publications have discussed and verified the capability of using cross-correlation data from microphone pairs to determine traffic parameters. However completed work did not include pattern extraction techniques for a fully automatic parameter extraction.

Road traffic noise, outdoor propagation and unsuccessful methods

The effects of outdoor sound propagation on audio signals measured at the microphone were considered in terms of the system proposed. The two signals measured by the slightly separated microphones have been subjected to the same level and type of outdoor sound propagation effects as they travelled from the vehicular sound source to the microphones. Secondly, the system is designed to measure the phased differences between the two signals, not determine source signal characteristics. As a result the propagation effects causing signal attenuation or distortion for the applicable range need not be taken into account, assuming there is no phase distortion. The effects of wind and temperature gradients under normal conditions may also be ignored, provided the distance between source and receiver is within a hundred meters. This is true for the system described in this thesis.

In order to consider the measured signal as a plane wave due to geometrical spreading, the distance between the microphones was chosen to be substantially less than the distance to the centre of the road. Precipitation, rain, snow, or fog have an insignificant effect on sound levels. Wet road surfaces alter the type of sound generated, but audio-based traffic monitoring systems described in the literature were found to perform well in adverse weather including rain.

An analysis of the sounds produced by road vehicles was performed prior to determining suitable vehicle detection approaches. The noise level and frequency spectrum of a vehicle is governed by a wide variety of parameters, from engine speed, vehicle velocity and road surface to environmental weather, background noise and receiver location. The use of the Doppler principle to estimate vehicle motion was described in research publications as demonstrating poor results. This is to be expected since a majority of the sound generated by a vehicle above 30 km/h is due to the tyre/road interaction which has little if any harmonic content. The frequency spectrum of a vehicle was found to be relatively flat wide-band noise that does not differ hugely within vehicle class or from one class of vehicle to another. For this reason it is difficult to classify a vehicle based on frequency spectrum alone.

It was decided not to develop a learning-based traffic monitoring system that requires the recognition of a sound as vehicular noise, for the following reasons. The sound generated by road vehicles is changing as technology advances and more modern vehicles are produced, and is likely to continue to do so in the future. Manufacturers have reduced the previously dominant engine noise, resulting in tyre/road noise becoming the predominant vehicular sound source. Currently manufacturers are working towards reducing tyre/road noise, which will once more alter the generated sound characteristics. Therefore, it is impossible to exactly define temporal-spectral vehicular noise characteristics, since these may change over time. As mentioned, there is a lack of distinction in the characteristics of sounds generated by vehicles, adding to the difficulty in distinguishing individual sources in traffic.

Early experiments were performed to test the use of a large range of existing audio features in detecting the presence of a vehicle. These features included the average zero-crossing rate, signal energy, spectral centroid and fundamental frequency. Audio features are typically used in sound classification and separation, where there is little noise and the temporal-spectral characteristics of the sound being examined are distinctive and distinguishable. They are generally not designed to be robust to uncontrolled outdoor environments or to detect sounds whose characteristics are often negligibly different to the background noise. It was found that sound amplitude was the only feature vector to change noticeably in the presence of a passing vehicle. It was concluded that the examined audio features are not suitable for traffic monitoring, so an audio feature-based approach was not considered further. A description of investigations involving audio features is included in Appendix A to justify this decision.

Source localization techniques

A microphone array needs to distinguish between multiple distributed vehicle sources as well as determine relevant motion characteristics of each vehicle. Source localization techniques determine the spatial location of a source based on multiple observations of the emitted sound signal. In sound source localization, the desired information is the position of the sound emitting source; the acoustical characteristics are largely irrelevant. A minimum of two spatially distributed sensors are required to determine the location of a source.

Beamforming is one possible localization technique. However, it has high computational requirements due to the large number of sensors and signal processing necessary. This prohibits its use in the majority of practical, real-time source locators. A further limitation is that the beamformer performance is directly dependent upon the physical size of the sensor array, and performance is suboptimal when using a small number of microphones. The objective of this research was to develop a simple and efficient traffic monitoring system, and consequently beamforming was determined not to be a suitable approach.

The signals received by microphones in an array due to an emitted sound are time-shifted versions of one another to a very good approximation. A TDOA approach uses this signal similarity to determine the inter-signal time delay. This is achieved by cross-correlating two microphone signals and determining the time delay by the distance of the maximum cross-correlation from the origin. Primarily because of their computational practicality and high performance, many passive localization systems are TDOA-based. Cross-correlation based TDOA is reported to have an inability to accommodate multi-source scenarios since these algorithms assume a single source model. However, cross-correlation experiments described in Chapter 8 demonstrate that multiple sound sources were successfully detected. It is better suited to vehicle tracking with a small microphone array and demonstrates reliable performance in adverse conditions. For these reasons, it was decided to use a TDOA-based localization approach as the basis for traffic monitoring in the work described here. Cubic spline interpolation was used to decrease the bin size and increase the bin resolution.

Cross-correlation via the frequency domain was faster and allows the possibility of emphasising the phase information in the data. The source signal auto-correlation component of the cross-correlation function has the detrimental effect of broadening the time-delay peak of interest. Multiple time delays can spread into one another, thereby making it impossible to distinguish delay times. Cross-correlation data obtained from traffic recordings demonstrated that the spreading effect impeded the accurate measurement of time delay. It was therefore necessary to apply some weighting function to reduce the effect of the source signal auto-correlation component. No prior knowledge of the source signal characteristics is available. Therefore, a general weighting function was used that flattens the magnitude of the frequency domain cross-power spectral density. As a result, the pattern created by a passing vehicle in the weighted cross-correlation array was more defined and distinguishable from background noise, despite the magnitude being lower.

A side effect was found in the application of weighting to the frequency-domain cross-spectral density. The flattened magnitude component approximates a DC signal. Consequently, the DC signal transforms to the time domain as a sinc function overlaid on the phase difference information containing the inter-microphone time delay. However, this was taken into account during the pattern extraction stage by weighting the central cross-correlation values less favourably.

Geometrical model and parameter evaluation

A moving source model was developed that mathematically describes the location of a sound source, based on inter-signal time delay and known microphone array geometry. One of the benefits of modelling the sound source behavior was the ability to perform calculations for a range of variables and parameters such as source velocity. In this manner, results were calculated to evaluate the trade-off between parameters and quantify the accuracy a particular set of values may achieve. This reduced the need for exhaustive measurements. Real data was compared against an accurate model to ascertain the vehicle characteristics.

The inter-microphone distance parameter m was found to be highly relevant, as it influences system accuracy and is a key parameter in dictating the shape of the moving source model. The further apart the microphones are, the greater the maximum measurable time delay τ_{max} will be, making it easier to distinguish different locations. However, m needs to be substantially less than D, which in turn should not be so great that the increasing sound attenuation reduces the accuracy in cross-correlation time delay measurement.

The resolution and number of measured source locations over time was determined by the sampling frequency and vehicle velocity, since a fast-moving vehicle will be changing location between the sampling times. The observed road length was calculated as 2 to 6 meters, depending on the geometry of the equipment at each recording location. For an observed road length of 2m, the number of measurements was approximately 30. Based on the sampling frequency and interpolation level used, the precision in measuring velocity was calculated to be within an accuracy of ± 5.88 km/h.

One purpose of the choice in window length was to ensure the spectral characteristics are reasonably stationary over the duration of the window, since stationarity is a requirement for the cross-correlation method implemented. However, an appropriate window size that achieved wide-sense signal stationarity could not be defined for the recorded audio traffic signals. This is due to the fact that all window sizes resulted in a large variation in statistical characteristics, making any stationarity assumption invalid. Nevertheless, both the cross-correlation sequence and Fourier transform methods performed as expected, despite the stationarity assumption not being satisfied.

Pattern extraction

Three different lists of reference data were manually generated - audio, video and cross-correlation ground truth. The cross-correlation ground truth was found to be more accurate than the audio signals and was used as the reference data in evaluating the automatic traffic monitoring systems. Three different systems were developed and tested: an acoustic amplitude-based approach and two cross-correlation methods designed to extract time-delay patterns via peak tracking and shape matching.

Acoustic amplitude-based vehicle detection

The acoustic amplitude-based approach detects local maxima in a smoothed, low-pass filtered version of the energy vector. It determines whether these local maxima represent vehicles based on a number of criteria, including checking the frequency spectrum.

One of the difficulties was in identifying whether a temporary increase in sound amplitude is due to a single noisy vehicle or a group of quiet vehicles in close proximity. A particularly loud vehicle or background noise was found to sometimes acoustically mask successive quiet vehicles. For these reasons, estimating the amount of vehicles present was found to be highly prone to errors and resulted in a number of missed vehicles. Since the background noise typically changes over time, an adaptive threshold is required to determine which sound amplitude peaks possibly represent vehicles. However, it is impossible to distinguish background noise without knowing the acoustical properties of the audio event to be detected. Therefore background noise that is similar to the audio event was falsely classified as an event. When the threshold value was too high vehicles are missed, but if the threshold is too low there is an excessive number of false detections.

The acoustic amplitude-based traffic monitoring had difficulties in accurately determining the amount of vehicles. It is clear from experimental results that an acoustic amplitude-based approach is an unreliable method for traffic monitoring, detecting approximately only 50% of vehicles. Furthermore, there were a number of false detections.

Cross-correlation peak tracking

The other two implemented traffic monitoring approaches detect and evaluate moving source behaviour based on evidence in a cross-correlation array. The first approach was designed to minimize the amount of data stored and analysed, by tracing the path of salient data and comparing the path behaviour to what is expected of a desired event. Only the larger peaks in each cross-correlation sequence are selected, the remainder of the array is discarded. For successive cross-correlation sequences over time, the propagation of each selected peak is analysed to form peak trails or paths. The resulting paths are analysed with reference to the expected moving source behaviour to produce a list of detected events. No assumption is made regarding the quantity or type of moving sources present in the data, to allow for the presence of multiple simultaneous sources.

Theoretically, the peak tracking approach should be highly efficient while accurately detecting vehicles. There is a reliance on highly accurate results at each stage, from the correct detection of all cross-correlation peaks and the linking of peaks to correct trails, to the correct matching of a model to a completed trail. From the moment that an error is introduced it propagates through the data to result in a false or missed detection or inaccurate measurement of the time of passage. The peak-tracking approach to vehicle velocity measurement returns unreliable velocity results that often differ greatly from the true value. It was realised that by its nature the peak-tracking approach is prone to selecting an incorrect model.

Cross-correlation shape matching

The second cross-correlation approach searches for regions of high correlation in the array that match the time-delay shape model of a passing vehicle. All array values within the region of a particular shape model are summed, in a similar manner to Hough shape detection. This is repeated for a range of model parameter values, with the results being mapped into the model parameter space. Clustering techniques are used to determine local maxima in the parameter space, which indicate a strong match between a particular model and the cross-correlation data. In this manner, passing vehicles and their parameter values are detected.

The first Hough method implemented in this project searched for rectangular regions of high correlation in the data since a moving sound source can be approximated with a line, particularly in the near-field scenario. A disadvantage was that the rectangular shape being sought was not the same as the modelled shape of a moving source. This means that even a perfect match based on a rectangle does not optimally represent actual source behaviour. Once the equations modelling a moving source were derived, it was possible to search for for a more precise shape than a rectangle. This improved vehicle detection accuracy and correct parameter estimation. Therefore the third traffic monitoring approach used the moving model in shape matching to detect vehicles and their parameters in the cross-correlation array.

Based on a substantial amount of traffic events, the shape matching pattern extraction approach was shown to be a highly accurate approach to detecting vehicles.

More than one unique velocity and reference time value was found to fit the cross-correlation data with high probability. This made it very difficult and at times impossible to define the most appropriate model, since a number of different models may be suitable. Therefore, the accuracy in determining the source velocity was found to be constrained by the sharpness of the cross-correlation array, resulting in a lower than expected velocity accuracy. However, a general indication of the vehicle speed was obtained and may be used provided precise speed measurements are not required.

When comparing the speeds of the traffic monitoring systems, the acoustic amplitude-based method was the fastest approach while the slowest method was the peak-tracking method. Both cross-correlation approaches were found to be sufficiently robust to background noise and were capable of distinguishing multiple vehicles.

9.2 Conclusions and future work

Experimental results based on the three developed traffic monitoring systems have demonstrated that the use of audio information to detect vehicular traffic is a viable option. There are numerous approaches to use audio information in monitoring traffic. However, not all of them succeed in consistently presenting reliable results. Two of the key challenges in audio traffic monitoring are distinguishing individual vehicles and extracting the vehicle information from the audio data so the traffic characteristics can be determined. Robustness to noise and source signal type is also desired in any system when detecting vehicles, particularly since vehicle noise differs from car to car and is changing constantly as technologies evolve. Cross-correlation based localization doesn't care what the signal characteristics are, only that there is sufficient correlation between the two microphone signals to determine the time delay. For this reason, the TDOA traffic monitoring approach is highly suitable. Since it is necessary to track how the cross-correlation time delay changes over time in order to track a moving source, a more robust method is to retain as much evidence as possible. This was achieved by implementing the shape matching pattern extraction method, which presented highly accurate results in vehicle detection.

For an objective evaluation of the audio-based traffic monitoring systems developed in this project, it is recommended to perform future experiments where existing technologies and the proposed system may be tested simultaneously on the same traffic data. A comprehensive analysis should test a large range of technologies over a long time period in a wide variety of weather and traffic conditions. Furthermore, a variety of road sizes, surfaces and traffic conditions are desired. Such a comprehensive series of tests would quantify the performance of each traffic sensing technology and determine the relative position in performance of the proposed audio-based traffic monitoring system.

In order to test the audio-based system over a long time duration, it is necessary for it to be permanently installed at an appropriate location. The entire system must be designed to be weather-proof, compact and efficient enough to operate with a limited power supply. Since the acoustical signals measured by the microphones must be processed by the permanently installed system, a hardware and real-time software implementation of the signal processing algorithms is required. This was not implemented during the course of this thesis.

The purpose of detecting vehicle characteristics is to provide this information to traffic management systems and road users. Therefore, once the system has successfully detected vehicle characteristics over time, this information must be communicated in some manner to a data collection point. The transmission of data may occur at regular time intervals regardless of the traffic behaviour, or occur once a particular amount of vehicles have passed the system. During quiet periods, it may be preferable for the system to hibernate in order to preserve battery life. In this case, a simple sound amplitude-based early warning system could be used to activate a more demanding cross-correlation based approach. It is recommended that data transmission technology be included in future audio-based systems, so they may be installed permanently at a road.

The traffic monitoring system developed in this project demonstrated a high performance for two lanes. When testing the system at a location with four lanes, it was found that the distance between the system and the outer two lanes was too great. This resulted in the signals from the outer vehicles being excessively attenuated and indistinguishable in the cross-correlation array. The restriction in multi-lane monitoring of vehicular traffic was found to be not necessarily the amount of lanes, but rather the distance between the system and lanes to be monitored. A recommended experiment is to place the traffic monitoring system in the median strip at the centre of a multi-lane road. If the microphones are omnidirectional and placed correctly, they are capable of detecting acoustic signals from lanes on both sides of the system.

In this manner, the ability of the system to detect vehicles in more than two lanes could be evaluated. A suitably safe and accessible location was not found during the time left to perform experiments prior to completion.

The performance strength and efficiency of the system developed in this research provides an incentive to install a series of audio-based traffic monitoring systems in different locations along a multi-lane road. With an ability to transmit traffic data, they could also be expanded to relay data to a traffic management system. Alternatively, these systems could share information to make decisions about mutually detected vehicles, enhancing the accuracy of results and physical area that is monitored. An audio distributed wireless sensor network draws its strength from the individual capabilities of each sensor system. Until a single autonomous sensor is established, it is premature to develop a network of sensors. Therefore, this research task is recommended as future work.

The traffic monitoring system developed in this research has demonstrated a wide range of velocity accuracies. This is due to the fact that a range of different models typically match the cross-correlation data in the shape matching approach, since the models are far more precise and narrow than the time-delay evidence in the data array. Therefore, to increase the velocity accuracy it is necessary to improve the match between the moving source model shape and the cross-correlation data. Two proposed solutions are to either increase the cross-correlation array resolution, or increase the width of the moving source model to more precisely match a measured passing vehicle. It is expected that the velocity accuracy would increase, since the number of matching models and therefore corresponding velocity parameter values would be significantly reduced.

It would be useful to develop a modified version of the moving source model that simulates the observed temporary splitting and merging of a measured sound source time-delay into two sources from each end of the vehicle when in close proximity to the microphone array. Such a time delay model is similar in shape to a hysteresis curve. With this model, vehicle characteristics may be more precisely determined. Since each end of a vehicle exhibits the same velocity, the accuracy in velocity estimation may be increased due to the doubling of evidence. It would not always be possible to match such a model to every vehicular time-delay pattern, since two distinct sources are not always detected in close proximity. Nevertheless, a hysteresis-type model may be a more accurate representation of the shape of the time-delay data.

9.3 Completion of research objectives

A number of different research objectives were specified during the early stages of this work, which are listed in Section 1.4. Each research goal was addressed and the outcome described over the course of the work.

The use of audio information to monitor traffic was tested, and found to be viable in a range of locations for the conditions tested. Some audio-based approaches proved to be unreliable, such as a sound amplitude-based method or using audio features to distinguish events. Other localization methods return reliable and accurate results during experiments.

A mathematical model was successfully derived that simulates the time-delay pattern created by a moving vehicle. Using this, simulations were performed for a range of system parameters. Experiments showed that the shape-matching cross-correlation system could reliably detect, distinguish and track multiple vehicles solely based on measured audio data. Relevant source characteristics were measured in this manner.

A fully automatic audio-based traffic monitoring system was developed that requires only two microphones. A powered, weatherproof hardware implementation is still required that may be permanently installed on a road. However, simulations and knowledge gained from this work form the first steps in designing such a system.

9.4 Prior publications

MPEG-1 Bitstreams Processing for Audio Content Analysis

Roman Jarina, Orla Duffner, Sean Marlow, Noel O'Connor and Noel Murphy, ISSC 2002 - Irish Signals and Systems Conference, Cork, Ireland, 25-26 June 2002

Road traffic monitoring using a two-microphone array

Orla Duffner, Noel O'Connor, Noel Murphy, Alan Smeaton and Sean Marlow, AES Convention 118th Convention 2005 May 28-31 Barcelona, Spain

Appendix A

Vehicle Event Classification with Audio Features

Known audio features were considered during the early stages of investigating the use of audio signals to passively monitor vehicular traffic. Early experiments were performed with a large range of audio features. They revealed that the only feature vectors to noticeably change were based on sound amplitude. The overall change in feature vectors was not sufficient to detect vehicles and certainly insufficient to make any decision about the vehicle behaviour. It was concluded that the audio features considered are not suitable for traffic monitoring. A description of the audio features is included in this appendix for reference purposes and because in research even a negative result can be useful to other researchers.

A.1 Feature-based Event Identification

An audio signal in its raw form is unwieldy and disguises much of the information being sought. Often this data can be reduced to a series of characteristics features relevant to the task in hand, which may be used to discern an audio event. Audio Classification is typically based on a two-step approach. The steps are feature extraction followed by some training system such as Hidden Markov Models or Neural Networks that makes classification decisions based on the features. There are a variety of well-established audio features used for a range of applications, from speaker segmentation and speech/music discrimination to music genre identification and rhythm detection. It was decided to select a suite of the most popular features and investigate their worth in gaining knowledge of traffic events. The event to be flagged is the presence or passage of a vehicle. Desirable information includes the vehicle direction, velocity and type. Furthermore, individual vehicles in a group should be distinguishable.

The relevance and value of each features must be determined to justify their inclusion, but how can one perform such a decision in a quantifiable manner? It was decided to extract and retain a significant portion of audio features and use Principle Component Analysis (PCA) to gain a deeper understanding of the driving forces behind the features. Therefore, the next step after feature extraction was to test the capabilities of a combination of all extracted features in monitoring vehicle behaviour. If the features were found to be beneficial, then by process of elimination the most relevant features could be pinpointed. If the combined features were not sufficient to monitor vehicle behaviour, then the whole approach must be re-evaluated.

Section A.2 presents an overview of the audio features used in this work and describes how the features were extracted. Data reduction using Principle Component Analysis is introduced in Section A.3, while Section A.4 describes experiments and results performed to test this method. Section A.5 finishes with the conclusions.

A.2 Audio Feature Extraction

There are a wide variety of well-documented time and frequency-domain audio features used for classification and other application [128, 57, 61, 115]. *Time-domain* processing methods involve the waveform of the signal directly. Some examples of representations of the signal in terms of time-domain measurements include average zero-crossing rate, signal energy and the auto-correlation function. They are attractive because the required processing is very simple and provide a useful basis for estimating important features of the signal. *Frequency-domain* techniques involve (either explicitly or implicitly) some form of spectrum representation, whereby the frequency spectrum is typically obtained from a Fast Fourier Transform (FFT) of a short segment of the audio waveform. As a result, these more computationally-intensive features can exploit knowledge of the frequency spectrum. Examples include spectral centroid, spectral roll-off and fundamental frequency. A variety of both time and frequency-domain features were extracted. Some of the features originate from the MPEG-7 audio low-level descriptor standard. Formally named *Multimedia*

Content Description Interface, MPEG-7 is a standard for describing the multimedia content data that supports some degree of interpretation of the information's meaning, which can be accessed by a device or computer code [130, 83, 109, 166]. The following section describes the features and the method of extraction.

Zero Crossing Rate

The zero crossing rate Z_n is a simple time-domain feature. The zero crossing rate is defined as the number of time-domain negative to positive crossings of a vector within a defined region of signal, divided by the number of samples of that region.

$$Z_n = \frac{1}{2N} \sum_{m=n-N+1}^{n} |sgn[x(m)] - sgn[x(m-1)]|, \tag{A.1}$$

where

$$\begin{aligned} sgn[x(n)] &= 1 \qquad x(m) \geqslant 0 \\ &= -1 \qquad x(m) < 0. \end{aligned}$$

Rough estimates of spectral properties can be obtained using a representation based on the short-time average zero-crossing rate. Figure A.3 illustrates the zero crossing rate for an audio file that contains sounds of 6 passing vehicles. The zero crossing rate does not change in any noticeable manner when a vehicle is present.

Subband Energy

The short-time energy of the signal proves a convenient representation that reflects the overall amplitude variations of an audio signal over time. The short-time energy of a discrete-time signal s at sample n can be simply defined as

$$E_n = \sum_{m=n-N+1}^{n} s^2(m). \tag{A.2}$$

That is, the short-time energy at sample n is simply the sum of squares of the N samples $n - N + 1$ through n. Sometimes it is useful to observe the change in energy of a specific frequency or frequency range within a broadband signal such as sound. For this reason the energies of a series of different frequency subbands were extracted as audio features. Table A.1 describes the frequency content of each subband.

Subband 1:	100 - 470 Hz	Subband 6:	2000 - 2370 Hz
Subband 2:	480 - 850 Hz	Subband 7:	2380 - 2750 Hz
Subband 3:	860 - 1230 Hz	Subband 8:	2760 - 3130 Hz
Subband 4:	1240 - 1610 Hz	Subband 9:	3140 - 3510 Hz
Subband 5:	1620 - 1990 Hz	Subband 10:	3520 - 3890 Hz

Table A.1: Subband energy frequency band

The fast-fourier transform of the windowed audio signal is obtained, giving the power frequency distribution. By isolating individual frequency subbands, the total energy for that frequency subband can be obtained. This operation was performed on 10 different frequency subbands of equal width between 100 and 4000 Hz. The FFT was 1/10th of a second or 100 milliseconds and the overlap was 1/50th second or 20 milliseconds. Furthermore, ratios between different frequency subbands were obtained as a second series of features. The calculated subband energy ratios are as follows:

$$\text{Energy ratio 1} = \frac{sb(1)}{\sum_{i=2}^{10} sb(i)} \qquad \text{Energy ratio 4} = \frac{\sum_{i=2}^{3} sb(i)}{\sum_{i=4}^{10} sb(i)} \tag{A.3}$$

$$\text{Energy ratio 2} = \frac{\sum_{i=1}^{2} sb(i)}{\sum_{i=3}^{10} sb(i)} \qquad \text{Energy ratio 5} = \frac{\sum_{i=1}^{5} sb(1)}{\sum_{i=6}^{10} sb(i)} \tag{A.4}$$

$$\text{Energy ratio 3} = \frac{sb(2)}{\sum_{i=2}^{10} sb(i)} \tag{A.5}$$

where $sb(i)$ is the ith subband.

Figures A.1 and A.2 illustrate energy subbands 1-5 and 6-10 respectively, for an audio file that contains sounds of 6 passing vehicles. It may be observed that the energy subband features change noticeably as a vehicle passes. Based on sound amplitude, these features increase in magnitude to reflect the increased noise due to a vehicle in close proximity. When there is no vehicle present, the magnitude drops to reflect the small level of background noise.

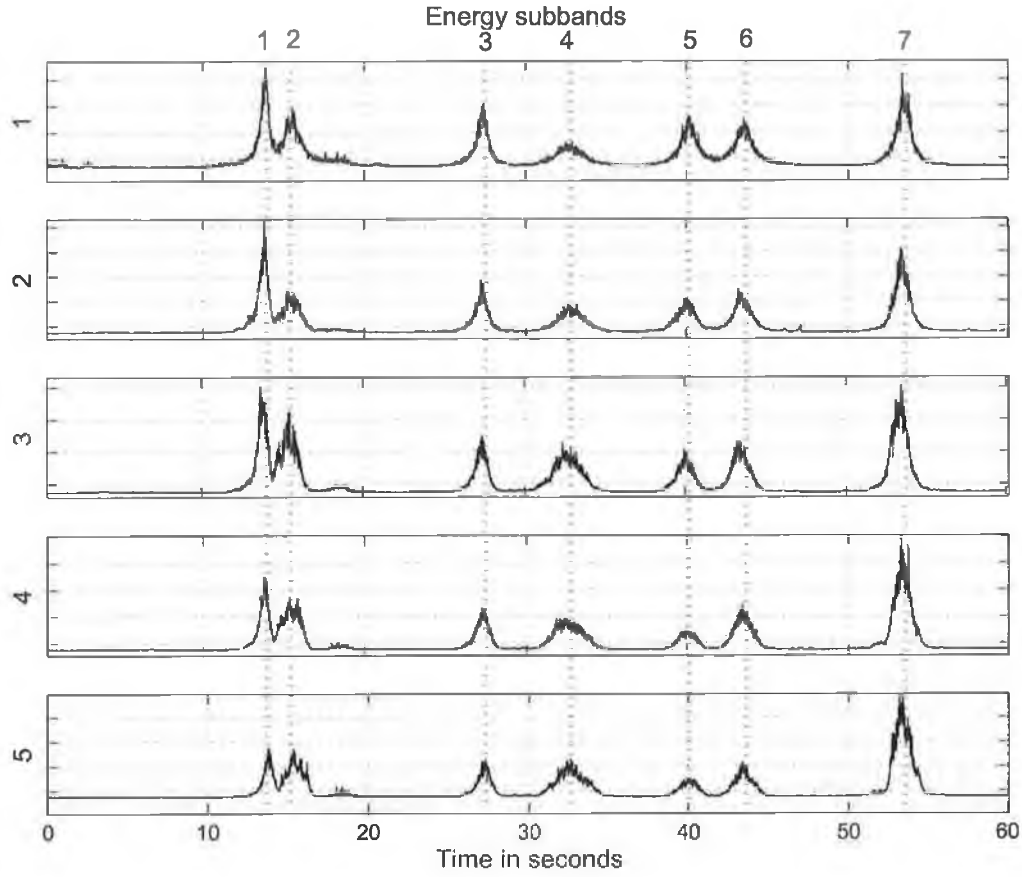


Figure A.1: Frequency spectrum energy ratio for subbands 1 to 5, described in Table A.1

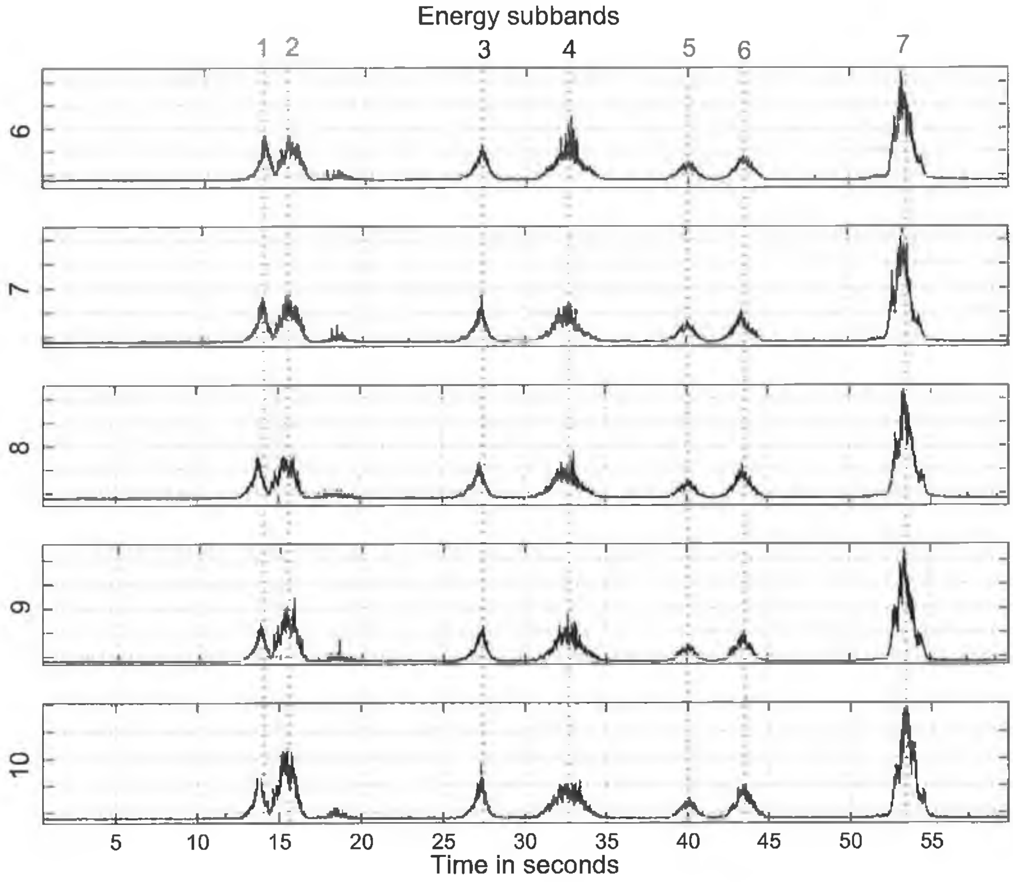


Figure A.2: Frequency spectrum energy ratio for subbands 6 to 10, described in Table A.1

Spectral Centroid

The spectral centroid is defined in the MPEG-7 low-level audio descriptor standard [130]. The MPEG-7 standard describes the centre of gravity of the log-frequency power spectrum, thereby indicating whether the power spectrum is dominated by low or high frequencies. The spectral centroid can be considered as the balancing point of the subband energy distribution. Equation A.6 defines the spectral centroid, c:

$$c = \frac{\sum_i log_2(\frac{f_i}{1000})p_i}{\sum p_i}. \tag{A.6}$$

The algorithm steps to obtain the centroid are summarized as follows:

- Calculate the DFT: 30ms segments of the signal are excised at 10ms intervals, a raised cosine window is applied, the window is zero-padded to the next power of two number of samples, and a FFT is performed;
- The power is calculated as the square magnitude of FFT coefficients;
- Samples below 62.5 Hz are replaced by a single sample, with power equal to their sum and a nominal frequency of 31.25 Hz;
- Frequencies of all samples are scaled to an octave scale anchored at 1kHz, and the spectrum centroid is calculated according to Equation A.6.

Figure A.3 illustrates the spectral centroid for an audio file that contains sounds of 6 passing vehicles. The spectral centroid displays a slight increase in magnitude as a vehicle passes. However, the increase is not very prominent or distinct. Furthermore, isolated increases also occur when a vehicle is not present.

Spectrum Flatness

The spectral flatness is an MPEG-7 feature describing the flatness properties of the short-term power spectrum within a given number of frequency bands. The spectral flatness expresses the deviation of the signal's power spectrum over frequency from a flat shape. A high deviation may indicate the presence of tonal components and may be used as a feature vector for robust matching between pairs of audio signals. The algorithm steps to obtain the centroid are summarized as follows:

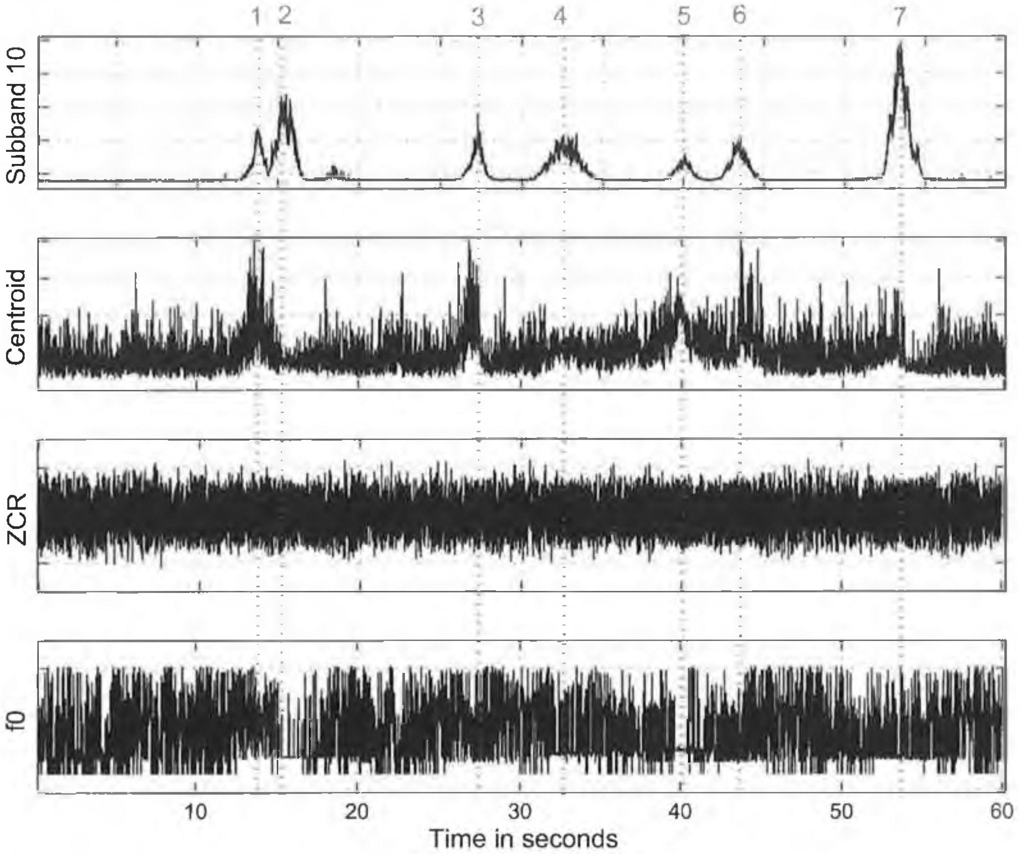


Figure A.3: Common audio features

- A spectrum analysis of the signal is performed as before, but with a hop size corresponding to the full window length;
- The flatness measure is calculated for a number of bands. These bands are defined by a partitioning of the frequency range from 300 Hz to 6300 Hz into bands of equal width. The spectral coefficients corresponding to each transition frequency are determined as described previously;
- For each band, the flatness measure is defined as the ratio of the maximum power spectrum coefficient and the mean of the power spectrum coefficients within the band. If no audio signal is present, a flatness measure of 1 is returned.

Spectrum Spread

Spectral spread, which is an MPEG-7 feature, describes the second moment of the log-frequency power spectrum and is defined as the RMS deviation of the log-frequency power spectrum with respect to its centre of gravity. The spectral spread is an economical descriptor of the shape of the power spectrum that indicates whether it is concentrated in the vicinity of its centroid or spread out over the spectrum. It enables differentiation between tone-like and noise-like sounds and is described in Equation (A.7).

$$s = \sqrt{\frac{\sum(log_2 \frac{f_i}{1000} - C)^2 p_i}{\sum p_i}} \tag{A.7}$$

The algorithm steps to obtain the centroid are summarized as follows:

- Calculate the power spectrum of the waveform and scale it to a log2 frequency scale. Samples below 62.5 Hz are grouped as before;
- Calculate the spectrum centroid as defined previously;
- Calculate the spectrum spread as the RMS deviation with respect to the centroid on an octave scale.

Fundamental Frequency

Most sounds are not pure tones with a single frequency but mixtures of different frequencies, the lowest of which is called the *fundamental frequency* (f_0). The estimation of the fundamental frequency is not an insignificant task, difficulties arise for a number of different reasons such as not all signals are periodic, the fundamental frequency may be changing over time, or signals may be contaminated with noise or signals with a different fundamental frequency. An existing MATLAB algorithm based on the subharmonic-to-harmonic ratio by Xuejing Sun [178, 179] was included to extract a fundamental frequency feature vector. The fundamental frequency vector is illustrated in Figure A.3 for an audio file that contains sounds of 6 passing vehicles. The fundamental frequency in Figure A.3 does not change in any noticeable manner when a vehicle is present.

Spectral roll-off

The spectral rolloff is the frequency under which 85% of the power distribution is concentrated. It is a measure of the amount of the right-skewedness of the power spectrum. In Equation (A.8), R is the rolloff frequency where $M[f]$ is the magnitude of the FFT at frequency f over N frequency bins:

$$\sum_{f=1}^{R} M[f] = 0.85 \times \sum_{f=1}^{N} M[f]. \tag{A.8}$$

A MATLAB program was written to create a feature vector based on the spectral roll-off for every signal section.

Mel-Frequency Cepstral Coefficients

Cepstrum analysis is a nonlinear signal processing technique with a variety of applications in areas such as speech and image processing. The cepstrum of a signal x is calculated by determining the natural logarithm of the magnitude of the Fourier transform of x, then obtaining the inverse Fourier transform of the resulting sequence, shown in Equation A.9:

$$\tilde{x} = \frac{1}{2\pi}\int log[X(e^{j\omega})]e^{j\omega n}d\omega. \qquad (A.9)$$

The audio signal is divided into short segments and passed through a Fast Fourier Transform to derive the harmonic power spectrum of each segment. That spectrum is then processed by a mel filter, which warps the spectrum according to the human auditory response as determined by decades of psychoacoustic research. Finally, the mel-filtered spectrum is subjected to a discrete cosine transform, which results in what is called a cepstrum consisting of multiple coefficients that represent the mel-adjusted shape of the original spectrum. Existing software in the MATLAB package *Voicebox* by Mike Brookes [30] was used to extract 13 mel-frequency cepstral coefficients for each Hamming windowed signal segment using the Discrete Cosine Transform and 32 filters in the filter-bank. MFCCs provide a compact representation of the spectral envelope such that most of the signal energy is concentrated in the first coefficients.

The first 6 mel-frequency cepstral coefficient audio features are illustrated in Figure A.4 for an audio file that contains sounds of 6 passing vehicles. They change slightly in the presence of a vehicle. However, it is difficult to distinguish the two vehicles that pass in close proximity. Furthermore, the change is not significant enough to facilitate the reliable detection of vehicles.

Linear Predictive Coding Coefficients

Linear prediction analysis determines a set of predictor coefficients α_k directly from the audio signal in such a manner as to obtain a good estimate of the signal properties [159]. These coefficients should be chosen as to minimize the error due to the difference between the actual and predicted signals. Because of the time-varying nature of the audio signal, the predictor coefficients must be estimated from short segments or windows of size m. Therefore, a more specific problem description would be that the predictor coefficients α_k should minimize the mean-squared prediction error E_n over a short segment of the audio waveform $s_n(m)$, as described in Equation A.10.

$$E_n = \sum_m e_n^2(m),$$

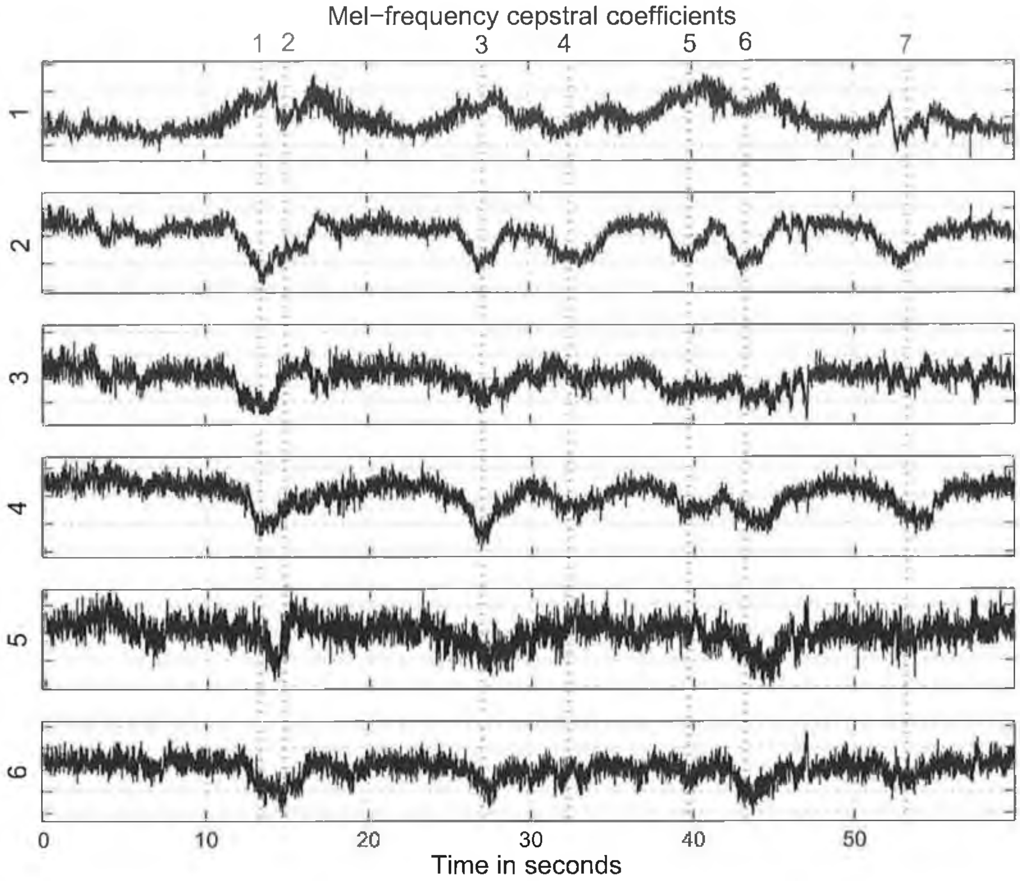


Figure A.4: Mel-frequency cepstral coefficient audio features

$$\begin{aligned} &= \sum_m [s_n(m) - \tilde{s_n}(m)]^2, \\ &= \sum_m \Big[s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k)\Big]^2. \end{aligned} \tag{A.10}$$

The resulting coefficients can then be used in a p^{th} order linear predictor of the audio signal. The values of α_k that minimize E_n can be obtained by setting $\frac{\partial E_n}{\partial \alpha_i} = 0$, $i = 1, 2, ..., p$, resulting in Equation A.11:

$$\sum_k^p \alpha_k \phi_n(i,k) = \phi_n(i,0) \qquad\qquad i = 1, 2, ..., p, \tag{A.11}$$

where

$$\phi_n(i,k) = \sum_m s_n(m-i) s_n(m-k). \tag{A.12}$$

Different techniques exist to model the audio signal s, such as the covariance, autocorrelation, lattice, spectral estimation, maximum likelihood and inner product methods. The autocorrelation technique was applied where normal equations that arise from the least-squares formulation were solved using the Levinson-Durbin recursion method. The MATLAB speech processing toolbox *Voicebox* [30] was used to obtain 21 linear prediction coefficients.

Statistical Features

A range of common statistical feature vectors were extracted from each audio section. These are summarized in Table A.2.

mean	average value in x_n	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
minimum	smallest value x_i in x_n	
maximum	largest value x_i in x_n	
median	centre value x_i in x_n	
standard deviation	deviation about the mean $\bar{x}$	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
variance	square of the standard deviation	

Table A.2: Extracted Statistical Features

and illustrated in Figure A.5 for an audio file that contains sounds of 6 passing vehicles. The statistical features in Figure A.5 are based on the sound amplitude.

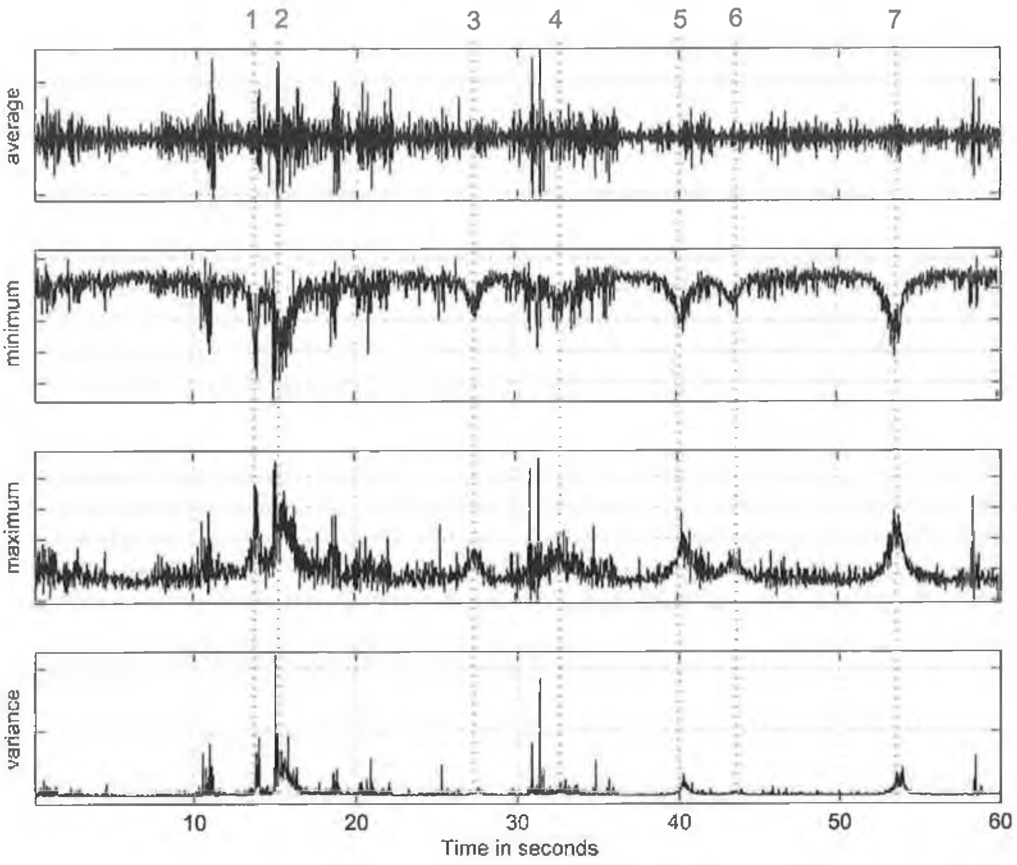


Figure A.5: Audio statistical Features

The amplitude maximum is observed as increasing in the presence of vehicles. The other statistical features vary to a lesser extent in the presence of vehicles.

A.3 Principle Component Analysis

One of the difficulties with extracting so many features is the problem of visualizing multi-dimensionality. A solution is to project the high-dimensional data onto a lower dimensional space using a classical approach known as *Principal Component Analysis* (PCA). It seeks a projection that best represents the data in a least-squares sense. The problem is to represent all the vectors in a set of n dimensional samples $x_1, ..., x_n$ by a single vector x_0. To be more specific, it is desired to find a vector x_0 such that the sum of the squared distances between x_0 and x_k is as small as possible. To find the best one-dimensional projection of the data, the data is projected onto a line through the sample mean in the direction of the eigenvector of the scatter matrix having the largest eigenvalue [55]. The principal components form an orthogonal basis for the space of the data and the sum of the variances of the first few principal components commonly exceeds 80% of the total variance of the original data.

Implementation

Principle Component Analysis uses standard statistical methods such as the covariance matrix, eigenvectors and eigenvalues. There are a series of steps involved in implementing PCA outlined as follows:

- Subtract the mean from each of the data dimensions, or $x_n - \bar{x}_n$ for each of the n-dimensions. This produces a data set whose mean is zero;

- The d-dimensional mean vector $\bar{x}$ and $d \times d$ covariance matrix are computed for the full data set;

- Compute the eigenvectors $(e_1, e_2, ..., e_d)$ and associated eigenvalues $(\lambda_1, \lambda_2, ..., \lambda_d)$ of the covariance matrix;

- Sort the eigenvectors and eigenvalues according to decreasing eigenvalue, sequentially naming the eigenvectors e_1 with eigenvalue λ_1, e_2 with eigenvalue λ_2 and so on. This sorts the components in order of significance;

- The first eigenvector with the largest eigenvalue is the principle component with the most significant relationship between the data dimensions;
- A feature vector is constructed by taking the first k eigenvectors and forming a matrix with k dimensions: $FeatureVector = (eig_1, eig_2, eig_3, ..., eig_p)$;
- The transposed feature vector is multiplied with the transposed original data set to generate the final data set.

Often there is just a few (or k) large eigenvalues, this implies that k is the inherent dimensionality of the subspace governing the signal while the remaining $d - k$ dimensions generally contain noise. Next a $d \times d$ matrix A is formed, whose columns consist of the k eigenvectors. The representation of data by principle components consists of projecting the data onto the k-dimensional subspace according to

$$x' = F_1(x) = A^t(x - \bar{x}). \tag{A.13}$$

Pre-multiplying the principle components by their transpose yields the identity matrix, confirming their orthogonality.

A.4 Experiments

A single omnidirectional microphone was placed adjacent to a two-lane bi-directional road to record audio data for analysis. This was digitized as a 16-bit wav signal with a sampling frequency of 44.1kHz. Due to the time-varying nature of the audio signal, it was necessary to analyse short overlapping segments of the signal at a time. The recorded signal was then processed as 30 millisecond windows of data with an overlap of 10 ms, during which relevant audio features were extracted and then analysed using purpose-built programs written in MATLAB code.

A sample of the extracted audio features are visible in Figures A.1 to A.4. The same audio recording is used to generate each audio feature, in which 7 vehicles pass the recording system in a 60-second time interval. The first 2 vehicles in the recording are in close proximity. The following 5 vehicles are distributed in time and space with time intervals of relative silence between vehicles. The largest change in audio features due to the presence of vehicles is evident in the features based on signal amplitude. The other features such as zero crossing rate and fundamental frequency display little or no change in the presence of vehicles.

A feature matrix was constructed with all the extracted feature vectors, after they had been normalized and re-scaled. Principle Component Analysis was performed on the matrix according to the steps described in Section A.3. The highest three principle components were multiplied with the original data and projected onto a 3-D video to observe their behaviour over time and gain a deeper understanding of the features variability as a vehicle passes.

A.5 Conclusion

Although there was some reaction to a vehicle passing, the pattern was unfortunately not predictable or reliable enough to be used as a rigorous vehicle tracking method. It was decided not to proceed with this method and pursue other options.

Appendix B

Cross-correlation array vehicle detection results

This appendix illustrates the cross-correlation array and presents selected vehicle detection results that are too large for the main body of the thesis.

In Section B.1 a series of cross-correlation images are shown. Gathered from various files, they illustrate the more challenging scenarios in vehicle detection. Events include simultaneous vehicles passing in opposite directions, multiple simultaneous sources, passing airplanes and trains. Section B.2 describes the type C data precision/recall results for shape matching pattern extraction.

B.1 Cross-correlation images

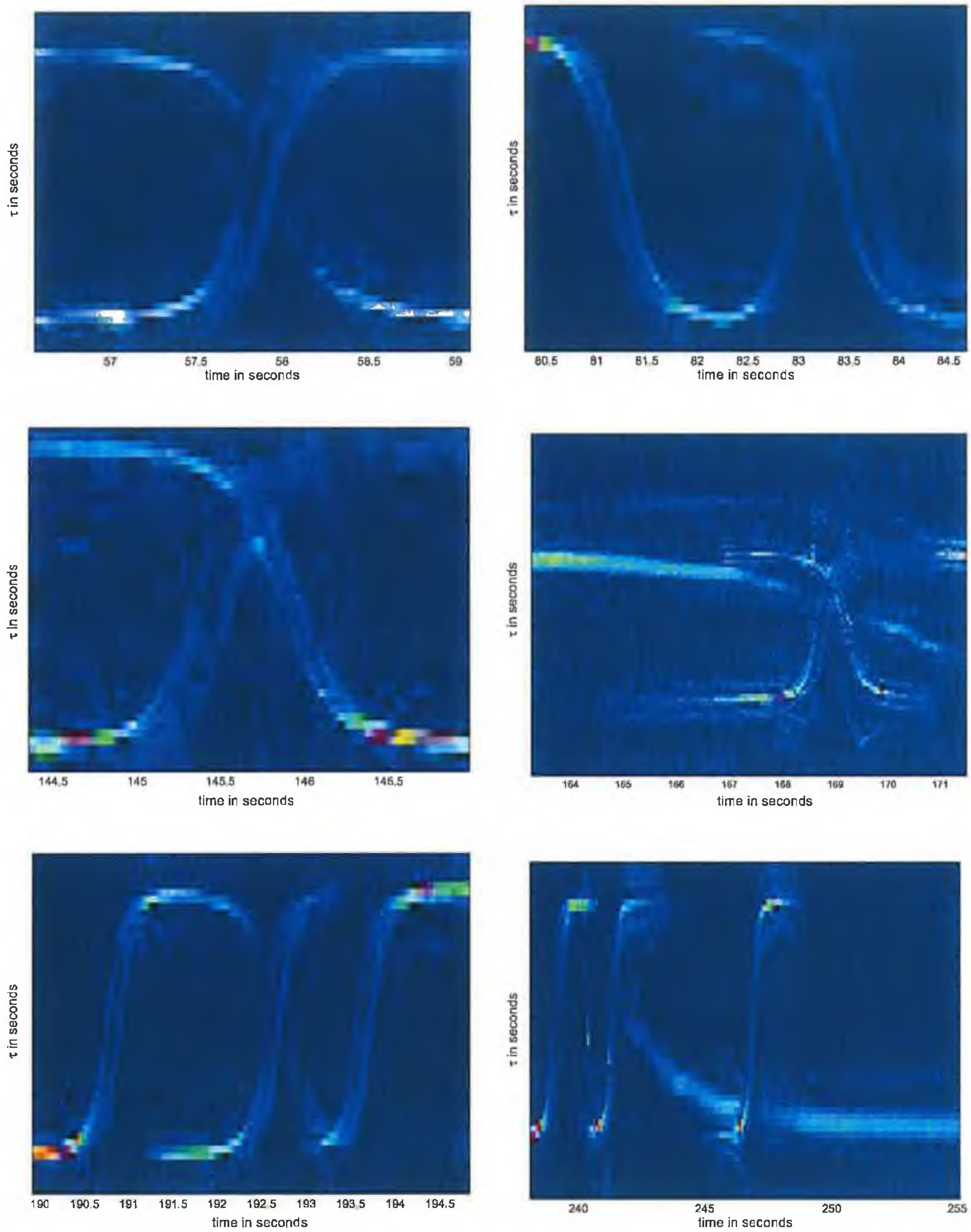


Figure B.1: Cross-correlation array segments for overlapping vehicles and airplanes, taken from files type A

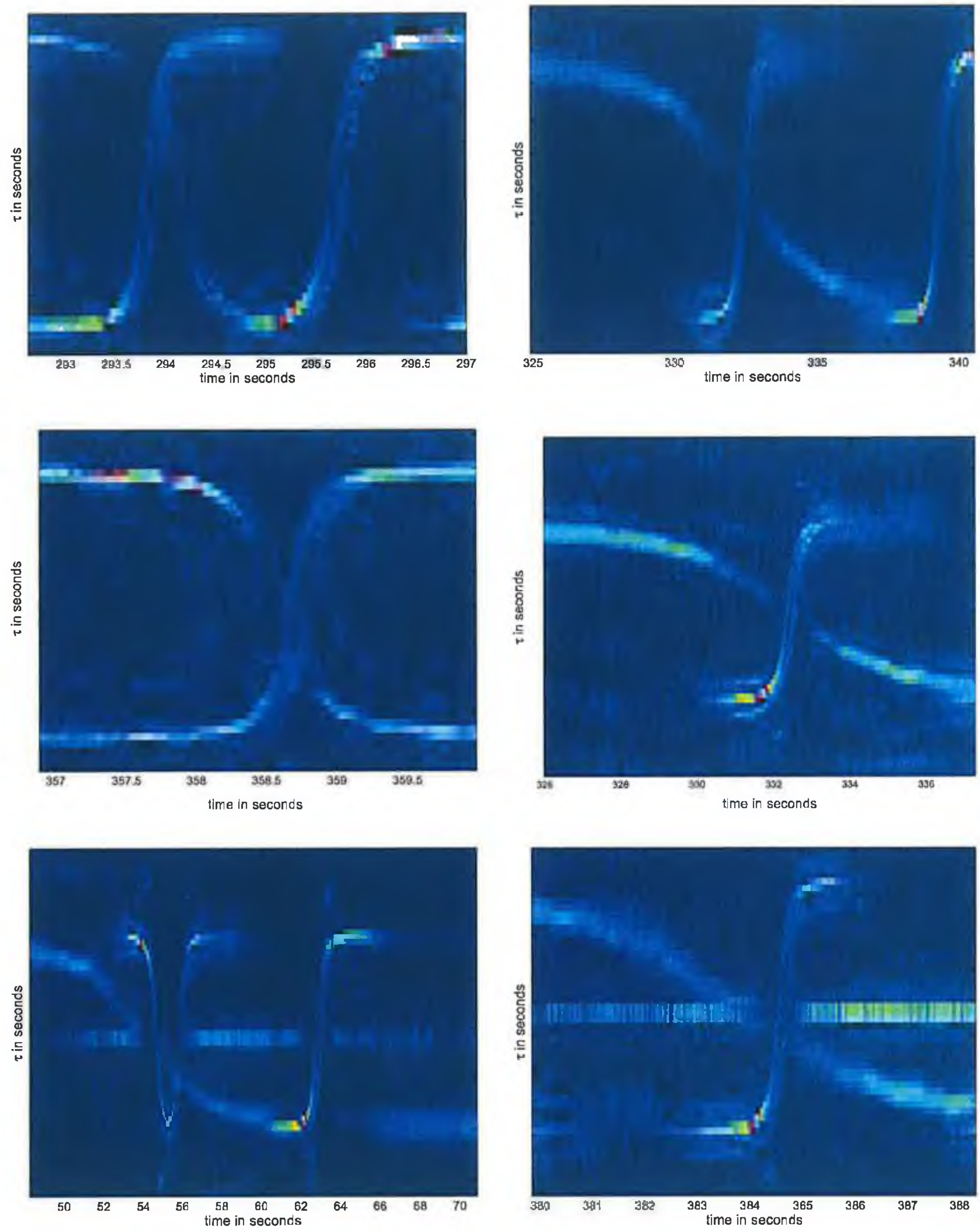


Figure B.2: Cross-correlation array segments for overlapping vehicles and airplanes, taken from files type A

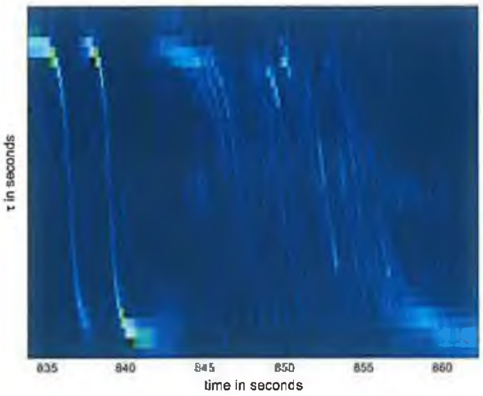


Figure B.3: Cross-correlation array segments for vehicles and a train, taken from files type B

B.2 Type C data precision/recall results for shape matching pattern extraction

As described in Section 7.3.2, a series of thresholds are required to optimize the selection of the most appropriate model in the parameter space of the shape matching pattern extraction method. These thresholds can be chosen with a view to optimizing precision or recall. Table B.1 presents the precision and recall values for different threshold values applied to the Type C recording. These values are illustrated in Figure B.4.

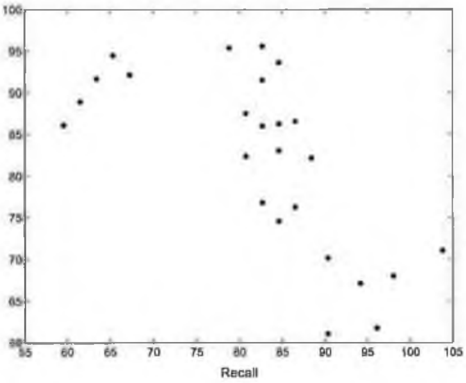


Figure B.4: Precision-Recall graph used to select the thresholds for optimizing results in file 12

Table B.1: Results for a range of shape matching thresholds applied to data recorded at location C

maximaLimit	dropEnough	tooClose	match	Hough	GT	Recall	Precision	
0.035	0.015	15	50	81	31	52	96.15	61.73
0.035	0.015	25	47	77	30	52	90.38	61.04
0.035	0.015	30	51	75	24	52	98.08	68
0.035	0.015	35	47	67	20	52	90.38	70.15
0.035	0.015	40	45	59	14	52	86.54	76.27
0.035	0.025	15	44	53	9	52	84.62	83.02
0.035	0.025	25	45	52	7	52	86.54	86.54
0.035	0.025	30	45	59	14	52	86.54	76.27
0.035	0.025	40	43	50	7	52	82.69	86
0.035	0.05	25	32	36	4	52	61.54	88.89
0.035	0.05	40	31	36	5	52	59.62	86.11
0.045	0.015	15	54	76	22	52	103.85	71.05
0.045	0.015	25	49	73	24	52	94.23	67.12
0.045	0.015	40	46	56	10	52	88.46	82.14
0.045	0.025	15	43	50	7	52	82.46	86
0.045	0.025	25	42	51	9	52	80.77	82.35
0.045	0.025	35	44	51	7	52	84.62	86.27
0.045	0.025	40	42	48	6	52	80.77	87.50
0.045	0.05	15	41	43	2	52	78.85	95.35
0.045	0.05	25	34	36	2	52	65.38	94.44
0.045	0.05	40	33	36	3	52	63.46	91.67
0.06	0.015	15	44	59	15	52	84.62	74.58
0.06	0.015	25	43	56	13	52	82.69	76.79
0.06	0.015	40	42	48	6	52	80.77	87.50
0.06	0.025	15	44	47	3	52	84.62	93.62
0.06	0.025	25	43	47	4	52	82.69	91.49
0.06	0.025	40	43	45	2	52	82.69	95.56
0.06	0.045	15	35	38	3	52	67.31	92.11
0.06	0.045	25	35	38	3	52	67.31	92.11
0.06	0.045	40	35	38	3	52	67.31	92.11
0.06	0.05	15	32	35	3	52	61.54	91.43

BIBLIOGRAPHY

[1] Acoustics - attenuation of sound during propagation outdoors, part 1: Calculation of the absorption of sound by the atmosphere. International Standard ISO 9613-1:1993(E), International Organisation for Standardization, Geneva, Switzerland, June 1993.

[2] Acoustics - specification of test tracks for the purpose of measuring noise emitted by road vehicles. International Standard ISO 10844:1994(E), International Organisation for Standardization, Geneva, Switzerland, September 1994.

[3] Acoustics - measurement of noise emitted by accelerating road vehicles - engineering method. International Standard ISO 362:1998(E), International Organisation for Standardization, Geneva, Switzerland, June 1998.

[4] High-speed ground transportation noise and vibration impact assessment. Technical Report 293630-1, U.S. Department of Transportation Federal Railroad Administration, Office of Railroad Development Washington, D.C. 20590, December 1998.

[5] Road traffic noise reducing devices - test method for determining the acoustic performance - normalized traffic noise spectrum. International Standard I.S. EN 1793-3, International Organisation for Standardization, Geneva, Switzerland, 1998.

[6] Encyclopedia Brittanica. 2005.

[7] Thushara D. Abhayapala and Hemant Bhatta. Coherent broadband source localization by modal space processing. In *10th International Conference on Telecommunications*, volume 2, pages 1617 – 1623, February - March 2003.

[8] D. G. Albert and J.G. Orcutt. Acoustic pulse propagation above grassland and snow: Comparison of theoretical and experimental waveforms. *Journal of the Acoustical Society of America*, 87:93–100, 1990.

[9] Donald G. Albert. Observations of acoustic surface waves in outdoor sound propagation. *Journal of the Acoustical Society of America*, 113(5):2495 – 2500, 2003.

[10] S. Applebaum. Adaptive arrays. *IEEE Transactions on Antennas and Propagation*, 24(5):585– 598, September 1976.

[11] G.P. Ashkar and J.W. Modestino. The contour extraction problem with biomedical applications. *Computer Graphics Image Processing*, 7:331–355, 1978.

[12] Keith Attenborough. Sound propagation close to the ground. *Annual Review of Fluid Mechanics*, 34:51 – 82, 2002.

[13] D. Aylor. Noise reduction by vegetation and ground. *Journal of the Acoustical Society of America*, 51:197205, 1959.

[14] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, 1999.

[15] Mike Baker. Lane width and its impact on road safety, capacity and the urban environment. *ITE Victoria*, September 2001.

[16] D. H. Ballard. Generalizing the hough transform to detect arbitrary shapes. Technical report, 1981.

[17] Yoshihide Ban, Hideki Banno, Kazuya Takeda, and Fumitada Itakura. Synthesis of car noise based on a composition of engine noise and friction noise. In *IEEE Proceedings*, pages 2105–2108, 2002.

[18] D. Bechler and K. Kroschel. Considering the second peak in the gcc function for multi-source tdoa estimation with a microphone array. In *8th International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 315–318, September 2003.

[19] Richard Bellmann. *Dynamic Programming.* Princeton University Press, 1957.

[20] Julius S. Bendat and Allan G. Piersol. *Engineering Applications of Correlation and Spectral Analysis.* John Wiley and Sons, 1980.

[21] Leo L. Beranek. *Acoustics.* Acoustical Society of America, 1993.

[22] E.F. Berliner, J. P. Kuhn, S. A. Rawson, and A. D. Whalen. Acoustic highway monitor. US Patent 6,021,364, Lucent Technologies Inc., 2000. Patent Date: Feb 1 2000, filed May 28, 1993.

[23] John W. Betz. Comparison of the deskewed short-time correlator and the maximum likelihood correlator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2):285–294, April 1984.

[24] D. Beymer, P. MacLaughlan, B. Coifman, and J. Malik. A real-time computer vision system for measuring traffic parameters. In *IEEE Computer Society Computer Vision and Pattern Recognition Conference*, pages 495–501, 1997.

[25] Stanley T. Birchfield and Daniel Kahn Gillmor. Acoustic source direction by hemisphere sampling. In *IEEE*, pages 3053–3056, 2001.

[26] Stanley T. Birchfield and Daniel Kahn Gillmor. Fast Bayesian acoustic localization. In *ICASSP*, volume 1, May 2002.

[27] D. E. Blumenfeld and G. H. Weiss. Attenuation effects in the propagation of traffic noise. *Journal of Sound and Vibration*, 9:103–106, 1975.

[28] D. E. Blumenfeld and G. H. Weiss. Curve fitting and probability distribution of acoustic noise from freely flowing traffic. *Journal of Sound and Vibration*, 12:111–114, 1978.

[29] M. Brandstein and D. Ward. *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer-Verlag, 2001.

[30] Mike Brookes. Voicebox. *MATLAB Central File Exchange*, (R11), February 2002.

[31] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:679–698, 1986.

[32] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *IEEE proceedings*, 57:2408–2418, August 1969.

[33] G. C. Carter. Coherence and time delay estimation. *IEEE Proceedings*, 1993.

[34] P. Castello, C. Coelho, E. Del Ninno, E. Ottaviani, and M. Zanini. Traffic monitoring in motorways by real-time number plate recognition. *IEEE International Conference on Image Analysis and Processing*, pages 1128 – 1131, September 1999.

[35] A. Chachich, A. Pau, A. Barber, K. Kennedy, E. Olenjniczak, J. Hackney, Q. Sun, and E. Mireles. Traffic sensor using a color vision method. In *SPIE Proceedings*, volume 2902, pages 156–165, Bellingham, Washington, 1996.

[36] C. H. Chen. *Nonlinear maximum entropy spectral analysis methods for signal recognition*. Pattern recognition & image processing. Research Studies Press, Chichester, England, 1982.

[37] Joe C. Chen, Kung Yao, and Ralph E. Hudson. Source localization and beamforming. *IEEE Signal Processing Magazine*, pages 30 – 39, March 2002.

[38] S. Chen, Z.P. Sun, and B. Bridge. Automatic traffic monitoring by intelligent sound detection. In *IEEE Conference on Intelligent Transportation System*, pages 171–176, November 1997.

[39] Shiping Chen, Ziping Sun, and Bryan Bridge. Traffic monitoring using digital sound field mapping. *IEEE Transactions on Vehicular Technology*, 50(6):1582–1589, 2001.

[40] Shu-Ching Chen, Mei-Ling Shyu, Srinivas Peeta, and Chengcui Zhang. Spatiotemporal vehicle tracking: the use of unsupervised learning-based segmentation and object tracking. *IEEE Robotics and Automation Magazine*, 12(1):50–58, March 2005.

[41] Xiuzhen Cheng, A. Thaeler, Guoliang Xue, and Dechang Chen. Tps: a time-based positioning scheme for outdoor wireless sensor networks. *IEEE Computer and Communications Societies Joint Conference (INFOCOM)*, 4:2685–2696, March 2004.

[42] C. F. Chien and Soroka. W. W. Sound propagation along an impedance plane. *Journal of the Acoustical Society of America*, 43:9–20, 1975.

[43] Ya-lun Chou. *Statistical analysis,: With business and economic applications.* Holt, Rinehart and Winston, 1969.

[44] Shanghai Maglev Transportation Development Co. http://www.smtdc.com/en/index.asp.

[45] Christophe Couvreur and Yoram Bresler. Doppler-based motion estimation for wideband sources from single passive sensor measurements. In *ICASSP*, volume 5, pages 3537–3540, April 1997.

[46] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. *23rd International Conference on Machine Learning (ICML)*, June 2006.

[47] Carl De Boor. *A Practical Guide to Splines.* Springer-Verlag, 2 edition, 2001.

[48] S.R. Deans. Hough transform from the radon transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3(2):185–188, March 1981.

[49] M.L. DeJong and B.E. Leonard. Demonstration of boyle's law on an air track. *American Journal of Physics*, 40(9):1342, September 1972.

[50] John H. Deller, John H.L. Hansen, and John G. Proakis. *Discrete-time processing of speech signals.* Wiley-Interscience, 2000.

[51] J. DiBiase, H. Silverman, and M. Brandstein. *Microphone Arrays: Signal Processing Techniques and Applications*, chapter Robust localization in reverberant rooms, pages 157–180. New York: Springer-Verlag, 2001.

[52] F. M. Dickey and K. S. Shanmugam. Optimum edge detection filter. *Applied Optics*, 16(1):145–148, 1977.

[53] F. M. Dommermuth and J. Schiller. Estimating the trajectory of an accelerationless aircraft by means of a stationary acoustic sensor. *Journal of the Acoustical Society of America*, 76(4):1114–1122, October 1984.

[54] R. J. Donato. Propagation of a spherical wave near a plane boundary with complex impedence. *Journal of the Acoustical Society of America*, 60:3–39, 1976.

[55] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification.* Wiley, Chichester, New York, 2 edition, 2001.

[56] R.O. Duda and P.E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):1115, 1972.

[57] Khaled El-Maleh, Mark Klein, Grace Petrucci, and Peter Kabal. Speech/music discrimination for multimedia applications. *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing*, 4:2445–2448, June 2000.

[58] Tony F. W. Embleton. Tutorial of sound propagation outdoors. *Journal of the Acoustical Society of America*, 100(1):31–48, July 1996.

[59] M. Fathy and M.Y. Siyal. A window-based image processing technique for quantitative and qualitative analysis of road traffic parameters. In *IEEE Transactions on Vehicular Technology*, volume 47, pages 1342–1349, November 1998.

[60] B. M. Favre and B. T. Gras. Noise emission of road vehicles: constitution of the acoustic signature. *Journal of Sound and Vibration*, 93(2):273–288, 1984.

[61] Jonathan Foote. Overview of audio information retrieval. *Multimedia Systems*, 7(1):2–10, January 1999.

[62] D. Forren. Traffic monitoring by tire noise. In *IEEE Conference on Intelligent Transportation System*, volume 3, pages 177–182, November 1998.

[63] J. Gajda, R. Sroka, M. Stencel, A. Wajda, and T. Zeglen. A vehicle classification based on inductive loop detectors. *IEEE Instrumentation and Measurement Technology Conference*, 1:460–464, May 2001.

[64] B. Gloyer, H.K. Aghajan, Kai-Yeung Siu, and T. Kailath. Vehicle detection and tracking for freeway traffic monitoring. *Signals, Systems and Computers Conference*, 2:970–974, October 1994.

[65] Y. Grenier. Time-dependent arma modeling of nonstationary signals. *IEEE Transactions on ASSP*, 31(4):899–911, August 1983.

[66] Scott M. Griebel and Michael S. Brandstein. Microphone array source localization using realizable delay vectors. *IEEE Workshop on the applications of Signal Processing to Audio and Acoustics*, pages 71–74, October 2001.

[67] S. Gupte, O. Masoud, R.F.K. Martin, and N.P. Papanikolopoulos. Detection and classification of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 3(1):37–47, March 2002.

[68] Don Halvorsen. Finger on the pulse - piezos on the rise. Technical report, Traffic Technology International, June/July 1999.

[69] R. C. Hansen. *Microwave Scanning Antennas*. Peninsular Publishing Company, June 1986.

[70] R. M. Haralick. Digital step edges from zero crossing of second directional derivatives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(1):56–68, 1984.

[71] Cyril M. Harris. Absorption of sound in air versus humidity and temperature. *Journal of the Acoustical Society of America*, 40(148), July 1966.

[72] Joseph C. Hassab, Brian W. Guimond, and Steven C. Nardone. Estimation of location and motion parameters of a moving source observed from a linear array. *Journal of the Acoustical Society of America*, 70(4):1054–1061, October 1981.

[73] M. Heckl, G. Hauk, and R. Wettschureck. Structure-borne sound and vibration from rail traffic. *Journal of Sound and Vibration*, 193(1):175–84, May 1996.

[74] Dietrich Heimann. Influence of meteorological parameters on outdoor noise propagation. *Acta Acustica*, 89, May/June 2003.

[75] Hsieh Hou and H. Andrews. Cubic splines for image interpolation and digital filtering. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(6):508–517, December 1978.

[76] M. Huadong, Wu Siegel, and P. Khosla. Vehicle sound signature recognition by frequency vector principal component analysis. In *IEEE Conference Proceedings of Instrumentation and Measurement Technology Conference (IMTC)*, volume 1, pages 429–434, May 1998.

[77] Y. Huang, J. Benesty, and G. W. Elko. Adaptive eigenvalue decomposition algorithm for realtime acoustic source localization system. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-97)*, pages 937–940, March 1999.

[78] Tarik M. Hussain, M. Baig Ahsen, Tarek N. Saadawi, and Samir A. Ahmed. Infrared pyroelectric sensor for detection of vehicular traffic using digital signal processing techniques. *IEEE Transactions on Vehicular Technology*, 44(3):683–689, 1995.

[79] J. Illingworth and J. Kittler. A survey of the hough transform. *Computer Vision, Graphics and Image Processing*, 44(1):87–116, 1988.

[80] U. Ingard. A review of the influence of meteorological conditions. *Journal of the Acoustical Society of America*, 25:405–411, 1953.

[81] M + P Raadgevende ingenieurs bv. www.mp.nl. Received during private correspondance with BMW, where the original graphs were generated by M+P in the Netherlands.

[82] Ramesh Jain, Rangachar Kasturi, and Brian G. Schunck. *Machine Vision*. McGraw-Hill, 1995.

[83] S. Jeannin and M. Bober. Description of core experiments for MPEG-7 motion/shape. MPEG-7, ISO/IEC/JTC1/SC29/WG11/MPEG99/N2690, Seoul, March 1999.

[84] D. R. Johnson and E. G. Saunders. The evaluation of noise from freely flowing traffic. *Journal of Sound and Vibration*, 7:287–309, 1968.

[85] Don H. Johnson and Dan. E. Dudgeon. *Array Signal Processing - Concepts and Techniques.* Prentice Hall, 1993.

[86] Hans G. Jonasson. Measurement and modelling of noise emission of road vehicles for use in prediction models. Technical Report 1999:35, Swedish National Testing and Research Institute, December 2000.

[87] K. P. Karmann and A. von Brandt. Moving object recognition using an adaptive background memory. *Proc. Time-varying image processing and moving object recognition*, 2, 1990.

[88] James H. Kell, Iris J. Fullerton, and Milton K. Mills. Traffic detector handbook. 2nd Edition FHWA-IP-90-002, U.S. Department of Transportation, FHWA, July 1990.

[89] David H. Kil. *Pattern recognition and prediction with applications to signal characterization.* Woodbury, N.Y: AIP Press, 1996. AIP series in modern acoustics and signal processing.

[90] Lawrence A. Klein. *Millimeter-wave and infrared multisensor design and signal processing.* Artech Houe, Norwood, MA, 1997.

[91] Lawrence A. Klein. *Sensor Technologies and Data Requirements for Intelligent Transportation Systems.* Artech Houe, Norwood, MA, 2001.

[92] Lawrence A. Klein and M. R. Kelley. Detection technology for ivhs. Volume 1 Final Report FHWA-RD-96-100, Hughes Aircraft Company, Turner-Fairbank Research Center, Federal Highway Administration Research and Development, U.S. Department of Transportation, Washington, D.C., 1996.

[93] C. H. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE TRans. Acous., Speech, Signal Processing*, 24:320–327, August 1976.

[94] H. Kobatake, Y. Inoue, T. Namai, and N. Hamba. Measurement of two-dimensional movement of traffic by image processing. In *ICASSP*, volume 12, pages 614–617, April 1987.

[95] S. Kok and P. Domingos. Learning the structure of markov logic networks. *Proceedings of 22nd International Conference on Machine Learning*, pages 441–448, 2005.

[96] J Kranig, E Ming, and C Jones. Field test of monitoring of urban vehicle operations using non-intrusive technologies. Technical Report FHWA-PL-97-018, Minnesota Department of Transportation, Minnesota Guidestar, St. Paul, MN and SRF Consulting Group, Minneapolis, MN, May 1997.

[97] Hamid Krim and Mats Viberg. Two decades of array signal processing research. *IEEE Signal Processing Magazine*, pages 67–94, July 1996.

[98] J. P. Kuhn. Detection performance of the smooth coherence transform (scot). In *ICASSP*, volume 3, pages 678–683, April 1978.

[99] J. P. Kuhn, B. C Bui, and G. J. Pieper. Acoustic sensor system for vehicle detection and multi-lane highway monitoring. US Patent 5,798,983, 1998. Patent Date: Aug 25 1998, Filed May 22 1997.

[100] U. J. Kurze. Statistics of road traffic noise. *Journal of Sound and Vibration*, 18:171–195, 1971.

[101] U. J. Kurze. Frequency curves of road traffic noise. *Journal of Sound and Vibration*, 33:171–185, 1974.

[102] Jet Propulsion Laboratory. Traffic surveillance and detection technology development, sensor development final report. Technical report, Federal Highway Administration, U.S. Department of Transportation, Washington, D. C., March 1997.

[103] C. Lamure. The annoyance due to road traffic noise, the mathematical modelling of such noise and the sound proofing of road vehicles. *Journal of Sound and Vibration*, 79(3):351–386, 1981.

[104] V.F. Leavers. *Shape Detection in Computer Vision Using the Hough Transform.* Springer-Verlag, 1992.

[105] V.F. Leavers. Which hough transform? *CVGIP: Image Understanding*, 58(2):250–264, 1993.

[106] H. Lee. A novel procedure for accessing the accuracy of hyperbolic multilateration systems. *IEEE Transactions on Aerospace Electronics*, AES-11:2–15, January 1975.

[107] Andre. L'Esperance, Yannick Gabillet, and Gilles A. Daigle. Outdoor sound propagation in the presence of atmospheric turbulence: experiments and theoretical analysis with the fast field program algorithm. *Journal of the Acoustical Society of America*, 98(1):570–579, July 1995.

[108] Xin Li, XiaoCao Yao, Yi. L. Murphey, Robert Karlsen, and Grant Gerhart. A real-time vehicle detection and tracking system in outdoor traffic scenes. In *17th International Conference on Pattern Recognition (ICPR'04)*, 2004.

[109] Adam Lindsay and Werner Kriechbaum. There's more than one way to hear it : Multiple representations of music in MPEG-7. *Journal of New Music Research*, 29(4):364-372, 1999.

[110] K.W. Lo and B.G. Ferguson. Broadband passive acoustic technique for target motion parameter estimation. *IEEE Transactions on Aerospace and Electronic Systems*, 36(1):163–175, January 2000.

[111] Roberto Lopez-Valcarce. Broadband analysis of a microphone array based road traffic speed estimator. In *IEEE Sensor Array and Multichannel Signal Processing Workshop*, July 2004.

[112] Roberto Lopez-Valcarce, D. Hurtado, Carlos Mosquera, and Fernando Perez-Gonzales. Bias analysis and removal of a microphone array based road traffic speed estimator. In *12th European Signal Processing Conference*, September 2004.

[113] Roberto Lopez-Valcarce, Carlos Mosquera, and Fernando Perez-Gonzales. Estimation of road vehicle speed using two omnidirectional microphones: a maximum likelihood approach. *EURASIP Booktitle of Applied Signal Processing*, 8(2):1059–1077, August 2004.

[114] Jianguang Lou, Tieniu Tan, Weiming Hu, Hao Yang, and S.J. Maybank. 3-d model-based vehicle tracking. *IEEE Transactions on Image Processing*, 14(10):1561–1569, October 2005.

[115] Lie Lu, Hong-Jiang Zhang, and Stan Z. Li. Content-based audio classification and segmentation by using support vector machines. *Multimedia Systems*, 8(6):482–492, April 2003.

[116] Art MacCarley. Advanced image sensing methods for traffic surveillance and detection. California PATH Research Report UCB-ITS-PRR-99-11, California Polytechnic State University San Luis Obispo, Institute of Transportation Studies, University of California, Berkeley, March 1999. ISSN 1055-1425.

[117] Art MacCarley, S. L. M. Hockaday, D. Need, and S. Taff. Evaluation of video image processing systems for traffic detection. Transportation research record 1360 - traffic operations, Transportation Research Board, Washington, D.C., 1992.

[118] Rufin Makarewicz. Air absorption of motor vehicle noise. *Journal of the Acoustical Society of America*, 80(2):561–568, August 1986.

[119] T. Malherbe and J. C. Bruyère. Ground effect on the acoustic directivity of a moving vehicle (influence du sol sur la directivité acoustique d'un véhicule en mouvement). Report, IRT-CERNE, 1981.

[120] C. Manning and H. Schutze. *Foundations of statistical natural language processing.* MIT Press, 1999.

[121] D. Marr and E. Hildreth. A theory of edge detection. Technical report, Royal Society of London B, 1980.

[122] A. Martelli. Edge detection using heuristic search methods. *Computer Graphics Image Processing*, 1:169–182, 1972.

[123] T. Matsuo, Y. Kaneko, and M. Matano. Introduction of intelligent vehicle detection sensors. In *IEEE Intelligent Transportation Systems Conference*, 1999.

[124] R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, August 1986.

[125] D. Middleton, D. Jasek, and R. Parker. Evaluation of the existing technologies for vehicle detection. Technical Report Research Project 0-1715, Research Report 1715-S, Texas Transportation Institue, College Station, TX, September 1999.

[126] Dan Middleton and Rick Parker. Initial evaluation of selected detectors to replace inductive loops on freeways. Technical Report Research Project 0-1439, Research Report 1439-7, Texas Transportation Institue, The Texas A&M University System, College Station, Texas 77843-3135, April 2000.

[127] Dan Middleton and Ricky Parker. Vehicle detector evaluation. Technical Report Research Project 0-2119, Research Report 2119-1, Texas Transportation Institue, The Texas A&M University System, College Station, Texas 77843-3135, October 2002.

[128] I. Mierswa and K. Morik. Automatic feature extraction for classifying audio data. *Machine Learning*, 58(2-3):127–49, February-March 2005.

[129] U. Montanari. On the optimal detection of curves in noisy pictures. *ACM Communications*, 14:335–345, 1971.

[130] ISO/IEC JTC1/SC29/WG11 N4980. MPEG-7 overview (version 8). July 2002.

[131] P. M. Nelson. *Transportation Noise Reference Book.* Butterworths, 1987.

[132] Y. Nishibe, N. Ohta, K. Tsukada, H. Yamadera, Y. Nonomura, K. Mohri, and T. Uchiyama. Sensing of passing vehicles using a lane marker on a road with built-in thin-film mi sensor and power source. *IEEE Transactions on Vehicular Technology*, 53(6):1827–1834, November 2004.

[133] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano. Localization of multiple sound sources based on a CSP analysis with a microphone array. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-97)*, pages 1053–1056, 2000.

[134] Amir Y. Nooralahiyan, Mark Dougherty, Denis McKeown, and Howard R. Kirby. A field trial of acoustic signature analysis for vehicle classification. *Transportation Research Part C: Emerging Technologies*, 5(3):165–177, 1997.

[135] Amir Y. Nooralahiyan and Howard R. Kirby. Vehicle classification by acoustic signature. *Mathematical Computer Modelling*, 27(9-11):205–214, 1998.

[136] S. Nordebo. *Robust broadband beamforming and digital filter design.* PhD thesis, Lule Univ. Technol., September 1995.

[137] H. Nyquist. Certain topics in telegraph transmission theory. *AIEE Transactions*, 47:617–644, April 1928.

[138] Institute of Transportation Engineers. *Traffic Detector handbook.* Number Draft 2. Federal Highway Administration, Washington D.C., July 2002.

[139] M. Ohta and Y. Mitani. A prediction method for road traffic noise generated from arbitrary non-poisson type traffic flow based on an approach equivalent to that for a standard posson type traffic flow. *Journal of Sound and Vibration*, 118(1):11–21, 1987.

[140] M. Ohta, S. Yamaguchi, and H. Iwashige. A statistical theory for road traffic noise based on the composition of component response waves and its experimental confirmation. *Journal of Sound and Vibration*, 52(4):587–601, 1977.

[141] R. A. Olson, R. L. Gustavson, R. J. Wangler, and R. E. McConnell. Active-infrared overhead vehicle sensor. *IEEE Transactions on Vehicular Technology*, 43(1):79–85, February 1994.

[142] M. Omologo and P. Svaizer. Use of the crosspower-spectrum phase in acoustic event localization. *IEEE Transactions on Speech and Audio Processing*, 5:288–292, May 1997.

[143] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *Discrete-time signal processing.* Prentice-Hall, 2 edition, 1995.

[144] Hough P. Machine analysis of bubble chamber pictures. *International Conference on High Energy Accelerators and Instrumentation, CERN*, pages 554–556, September 1959.

[145] Svaizer P., M. Matassoni, and M. Omologo. Acoustic source location in a three-dimensional space using crosspower spectrum phase. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-97)*, pages 231–234, April 1997.

[146] Sang Jin Park, Tae Yong Kim, Sung Min Kang, and Kyung Heon Koo. A novel signal processing technique for vehicle detection radar. *Microwave Symposium Digest*, 1:607 – 610, June 2003.

[147] A. Paulraj and T. Kailath. Direction of arrival estimation by eigenstructure methods with imperfect spatial coherence of wave fronts. *Journal of the Acoustical Society of America*, 83(3):1034–1040, March 1988.

[148] Li Peihong, Ding Liya, and Liu Jilin. A video-based traffic information extraction system. pages 528–532, 2003.

[149] Aron Pentek, James B. Kadtke, and Ron K. Lennartsson. Acoustic discrimination between aircraft and land vehicles using nonlinear dynamical signal models. In *International Symposium on Signal Processing and its Applications (ISSPA)*, volume 2, pages 687–690, August 2001.

[150] F. Pérez-Gonzáles, R. López-Valcarce, and C. Mosquera. Road vehicle speed estimation from a two-microphone array. In *IEEE ICASSP*, pages 1321–1324, 2002.

[151] Tien Pham and Manfai Fong. Real-time implementation of music for wideband acoustic detection and tracking. *SPIE AeroSence: Automatic Target Recognition VII*, April 1997.

[152] Tien Pham and Brian M. Sadler. Wideband array processing algorithms for acoustic tracking of ground vehicles. In *U.S. Army Research Laboratory.*

[153] L. S. Pontriagin. *The mathematical Theory of Optimal Processes.* Interscience, 1962.

[154] L. S. Pontryagin. *Optimal Control and Differential Games.* American Mathematical Society, 1990.

[155] G. Porges. *Applied acoustics.* London : Edward Arnold, 1977.

[156] John G. Proakis and Dimitris G. Manolakis. *Digital Signal Processing.* Prentice-Hall International, 3 edition, 1996.

[157] Z. Qiu, D. An, D. Yao, D. Zhou, and B. Ran. An adaptive kalman predictor applied to tracking vehicles in the traffic monitoring system. *IEEE Intelligent Vehicles Symposium*, pages 230–235, June 2005.

[158] Thomas F. Quatieri. *Discrete-time speech signal processing.* Prentice Hall Signal Processing Series, 2002.

[159] L. R. Rabiner and Schafer R. W. *Digital Processing of Speech signals.* Prentice-Hall, 1978.

[160] J Radon. über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten. *Berichte Sächsische Akademie der Wissenschaften*, 69:262277, 1917.

[161] V. Raghavan, P. Bollmann, and G. S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, 7:205–229, 1989.

[162] S. S. Reddi. Multiple source location - a digital approach. *IEEE Transactions on Aerospace Electronic Systems*, AES-15:95–105, January 1979.

[163] E. J. Richards and D. J. Mead. *Noise and acoustic fatigue in aeronautics.* Wiley, London, 1968. p.168.

[164] Thomas D. Rossing, F. Richard Moore, and Paul A. Wheeler. *The science of sound.* Addison Wesley, 3rd edition, 2002.

[165] I. Rudnick. Propagation of an acoustic wave along a boundary. *Journal of the Acoustical Society of America*, 19:348–356, 1947.

[166] P. Salembier and J.R. Smith. MPEG-7 multimedia description schemes. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):748–759, June 2001.

[167] U. Sandberg. Abatement of traffic, vehicle, and tire/road noise - the global perspective. *Noise Control Engineering Journal*, 49(4):170 – 181, 2001.

[168] Ulf Sandberg. Six decades of vehicle noise abatement - but what happened to the tyres? In *Acoustics 84*, pages 231–238, Institute of Acousitcs, U.K., 1984.

[169] Ulf Sandberg and Jerzy A. Ejsmont. *Tyre Road Noise Reference Book.* INFORMEX, Harg, SE-59040 Kisa, Sweden, 2002.

[170] K. Scarbrough, N. Ahmed, and D. H. Youn. On the scot and roth algorithms for time delay estimation. In *ICASSP*, volume 7, pages 371–374, May 1982.

[171] Kent Scarbrough, Nasir Ahmed, and G. Clifford Carter. An experimental comparison of the cross correlation and scot techniques for time delay estimation. In *ICASSP*, volume 5, pages 807–810, April 1982.

[172] R. Schmidt. A new approach to geomety of range difference location. *IEEE Transactions on Aerospace Electronics*, AES-8:821–835, November 1972.

[173] R. O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas Propagation*, 34:276–280, March 1986.

[174] C. Setchell and E.L. Dagless. Vision-based road-traffic monitoring sensor. *IEEE Proceedings on Vision, Image and Signal Processing*, 148(1):78–84, February 2001.

[175] H. Silverman, W. Patterson, J. Flanagen, and D. Rabinkin. A digital processing systm for source location and sound capture by large microphone arrays. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-97)*, pages 251–254, April 1997.

[176] D. Sturim, M. Brandstein, and H. Silverman. Tracking multiple talkers using microphone-array measurements. *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP-97)*, pages 371–374, April 1997.

[177] Thomas M. Sullivan. *Multi-microphone correlation-based processing for robust automatic speech recognition.* PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, August 1996.

[178] Xuejing Sun. A pitch determination algorithm based on subharmonic-to-harmonic ratio. *6th International Conference of Spoken Language Processing*, 4:676–679, April 2000.

[179] Xuejing Sun. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-97)*, 1:333–336, 2002.

[180] Y. Sun, B.G. Stewart, and I.J. Kemp. Alternative cross-correlation techniques for location estimation of PD from RF signals. *Universities Power Engineering Conference*, 1:143–148, September 2004.

[181] K. Takagi, K. Hiramatsu, and Y. Yamamoto. Investigations on road traffic noise based on an exponentially distributed vehicles model - single line flow of vehicles with same acoustic power. *Journal of Sound and Vibration*, 36:417–431, 1974.

[182] H.L. Tan, S.B. Gelfand, and E.J. Delp. A cost minimization approach to edge detection using simulated annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(1):3–18, 1992.

[183] D.W. Tindall. Application of neural network techniques to automatic licence plate recognition. *European Convention on Security and Detection*, pages 81–85, May 1995.

[184] Scott E Umbaugh. *Computer Imaging: Digital Image Analysis and Processing*. CRC Press, 2005.

[185] Robert J. Urick. *Principles of Underwater Sound*. Peninsular Publishing Company, August 1996.

[186] M.B. van Leeuwen and F.C. Groen. Vehicle detection with a mobile camera: spotting midrange, distant and passing cars. *IEEE Robotics and Automation Magazine*, 12(1):37–43, March 2005.

[187] van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann Ltd, March 1981.

[188] Harry L. Van Trees. *Optimal Array Processing*. Wiley, New York, 2002.

[189] B. Van Veen and K. M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, pages 4–24, April 1988.

[190] H. Wang and P. Chu. Voice source localization for automatic camera pointing system in videoconferencing. *IEEE Transactions on Acoustics, Speecn and Signal Processing*, pages 187–190, April 1997.

[191] Paul E. Waters. A review of road traffic noise. Technical Report LR 357, Road Research laboratory Report, Crowthorne England, 1970.

[192] A.J. Weiss. On the accuracy of a cellular location system based on RSS measurements. *IEEE Vehicular Technology Transactions*, 52(6):1508–1518, November 2003.

[193] Frederick A. White, editor. *Our acoustic environment.* Wiley-Interscience, 1975.

[194] F.M. Wiener and D.D. Keast. Experimental study of the propagation of sound over ground. *Journal of the Acoustical Society of America*, 31:724, 1959.

[195] D. Williams and W. Tempest. Noise in heavy goods vehicles. *Journal of Sound and Vibration*, 43(1):97–107, 1975.

[196] Guorong Xuan, Yang Xiao, Xiaoguang Yang, Jidong Chen, Chengyun Yang, and Ruhua Zhang. Traffic measuring and authentification on digital video. In *IEEE*, pages 725–729, 2003.

[197] Y. Yanamura, M. Goto, D. Nishiyama, M. Soga, H. Nakatani, and H. Saji. Extraction and tracking of the license plate using hough transform and voted block matching. *IEEE Intelligent Vehicles Symposium*, pages 243–246, June 2003.

[198] Tatsuro Yano, Takeshi Tsujimura, and Koichi Yoshida. Vehicle identification technique using active laser radar system. In *IEEE Proceedings on Multisensor Fusion and Integration for Intelligent Systems, MFI2003*, pages 275 – 280, 2003.

[199] N. Zeng and J. D. Crisman. Evaluation of color categotization for representing vehicle colors. In *SPIE Proceedings*, volume 2902, pages 148–155, Bellingham, Washington, 1996.

[200] C. Zhang, S.C. Chen, M.L. Shyu, and S. Peeta. Adaptive background learning for vehicle detection and spatio-temporal tracking. *International Conference on Information, Communications and Signal Processing*, 2:797–801, December 2003.

[201] Yufang Zhang, P. Shi, E.G. Jones, and Q. Zhu. Robust background image generation and vehicle 3d detection and tracking. *IEEE Conference on Intelligent Transportation Systems*, pages 12–16, October 2004.