Constant Power - Continuously Variable Transmission (CP-CVT): Optimisation and Simulation

A Thesis Submission for the degree of
Doctor of Philosophy

by
Colin Alexander Bell

MECHANICAL ENGINEERING
SCHOOL OF ENGINEERING AND DESIGN

Brunel University

2011

I. ABSTRACT

A novel continuously variable transmission has previously been designed that is capable of addressing a number of concerns within the automotive industry such as reduced emissions. At the commencement of this research, the design was in the early stages of development and little attempt had been previously made to optimise the design to meet specific measurable targets. This thesis utilises and modifies several design approaches to take the design from the concept stage to a usable product. Several optimisation techniques are adapted and created to analyse the CVT from both a design and tribological prospective. A specially designed optimisation algorithm has been created that is capable of quickly improving each critical component dimension in parallel to fulfil multiple objectives. This algorithm can be easily adapted for alternative applications and objectives.

The validity of the optimised design is demonstrated through a simulation-tool that has been created in order to model the behaviour of the CVT in a real automotive environment using multiple fundamental theories and models including tire friction and traction behaviour. This powerful simulation tool is capable of predicting transmission and vehicular behaviour, and demonstrates a very good correlation with real-world data. A design critique is then performed that assesses the current state of the CVT design, and looks to address some of the concerns that have been found through the various methods used. A specific prototype design is also presented, based on the optimisation techniques developed, although the actual creation of a prototype is not presented here.

Additional complementary research looks at the accuracy of the tire friction models through the use of a specially design tire friction test rig. Furthermore, a monitoring system is proposed for this particular CVT design (and similar) that is capable of continuously checking the contact film thickness between adjacent elements to ensure that there is sufficient lubricant to avoid metal-on-metal contact. The system, which is based around capacitance, requires the knowledge of the behaviour of the lubricant's permittivity at increased pressure. This behaviour is studied through the use of a specially-designed experimental test rig.

II. CONTENTS

III. SELECTED PRESENTATIONS AND PUBLISHED WORK

Bell, C., Glovnea, R.P. and Mares, C., 2008. "Concept Design for Transmission Systems", *Proceedings of International Multi-Conference on Engineering and Technological Innovation: IMETI*, June-July 2008, Orlando, FL, USA.

Bell, C. and Glovnea, R.P., 2009. "Modelling and Simulation of a Novel Toroidal-Type CVT", *Proc. Of MPT2009, JSME International Conference on Motion Power Transmissions*, May 2009, Sendai, Japan.

Bell, C. and Glovnea, R.. 2009. "Vehicular Simulation of a Toroidal-Type CVT", *The 2nd SED Research Conference*, 22nd-24th June 2009, Brunel University, Uxbridge, UK.

Bell, C.A., and Glovnea, R.P., 2010. "Tribological Optimisation of a Toroidal-Type CVT", *37th Leeds-Lyon Symposium on Tribology*, 7-10th September 2010, Leeds, UK.

Bell, C.A., and Glovnea, R.P., 2011. "Tribological Efficiency Optimisation of a Toroidal-Type CVT", *Proceedings of the IMechE, Part J, Journal of Engineering Tribology*. (accepted for publication, awating publication details).

Bell, C.A., Mares, C. and Glovnea, R.P., 2011. "Concept design optimisation for Continuously Variable Transmissions", *Int. J. Mechatronics and Manufacturing Systems*, Vol. 4. No. 1. 2011.

Glovnea, R. and Bell, C. 2009. "Design for fuel efficiency of an automotive toroidal CVT", *IOP: Tribology for Efficiency and Energy Saving*, 17th February 2010, Institute of Physics, London, UK

Morita, T., Imayoshi, Y., Bell, C., Glovnea, R. and Sugimura, J., 2010. "Experimental study of rubber traction with concrete model surfaces", *International Tribology Congress - ASIATRIB 2010*, 5th-9th December 2010, Perth, Australia.

Nagata, Y., Furtuna, M., Bell, C. and Glovnea, R., 2008. "Evaluation of Electric Permittivity of Lubricating Oils in EHD Conditions", *Proc. of 2nd International Conference on Advanced Tribology*, December 2008, Singapore

IV. ACKNOWLEDGEMENTS

Firstly, I would like to express my thanks to my supervisors Dr R. Glovnea and Dr C. Mares for their continued support, advice and guidance throughout my research. Additionally, I would like to thank Dr M. Furtuna for his continued help throughout.

Furthermore, I would like to express my eternal gratitude to the Thomas Gerald Gray Charitable Trust for their financial support throughout and for giving me the opportunity to pursue my own research. Without their funding this research would not have been possible.

I would also like to thank my partner, family and friends for their patience and encouragement during the PhD programme. A special thanks also goes to my brother, parents and grandparents, who have supported and encouraged any endeavour I have undertaken in my life and shown me that anything is possible.

For Mum and Dad

V. NOMENCLATURE

Lower Case

a	Semi-width of contact ellipse along longer axis
b	Semi-width of contact ellipse along shorter axis
c	Effective contact semi-width
d	Indicative diameter of CVT components
d_{m}	Bearing pitch diameter used in bearing-loss calculations
f	Frequency in capacitance calculations
h_c	Central contact film thickness
i	Transmission ratio, or normalised transmission ratio range
i_f	Final or differential transmission ratio
k	Geometry factor of bearing, or spring stiffness coefficient
l	Indicative length of CVT components, also used to indicate ball-screw lead length
m	Indicative mass of components, also used to indicate slope of traction curve
m_v	Vehicular mass
p_m	Mean Hertzian pressure
p_0	Peak Hertzian pressure
$p.v.$	Power-variation coefficient (correlation coefficient)
r_0	Distance from centre shaft to centre of radius of curvature of toroidal disc
s	Scoring function of each technical requirement
t	Time
u	Linear speed, with subscript indicating precise usage
v	Vehicle velocity in tire-friction calculations
w	Relative weighting factor of each technical requirement
x	Toroidal disc position relative to unloaded point
$\boldsymbol{x}$	Dimensional vector set of component dimensions β, γ, R, R_1, r_0
y	Offset of the location of the maximum load point in tire friction model

Upper Case

A	Referring to contact between ball element and toroidal input disc
A	Area of immersed surface, also frontal vehicle area in drag calculations
B	Referring to contact between ball element and conical input disc
C	Referring to contact between ball element and conical output disc
C	Capacitance
C_D	Coefficient of drag in vehicle motion calculations
C_M	Moment coefficient in churning loss calculations
C_R	Coefficient of rolling resistance in vehicle motion calculations
D	Diameter of roller elements used in bearing-loss calculations
$\overline{E}$	Reduced elastic modules
E	Elastic/Young's modulus of contacting material
F	Linear force, with subscript indicating precise usage

F_{dr}	Linear force due to drag in vehicle motion calculations
F_{rr}	Linear force due to rolling resistance in vehicle motion calculations
F_{tr}	Linear tractive force acting in primary traction direction
F_z	Linear driving force of vehicle in vehicle motion calculations
F_1	Variable used in Hertzian contact calculations
G	Average fluid shear modulus, or geometry factor of bearing
J	Moment of inertia
L	Chord length of immersion line, also contact patch length in tire friction
M	Operational torque of roller bearing
N_A	Normal force acting between toroidal input disc and ball elements
N_B	Normal force acting between conical input discs and ball elements
N_C	Normal force acting between conical output discs and ball elements
N_w	Total normal force acting between tires and road surface of vehicle
P	Power, also pitch length of macroscopic asperities for tire friction calculations
P_{new}	New point probability in particle swarm optimisation
R	Radius of ball elements, and other general radius dimensions
R_1	Radius of curvature of toroidal disc
R_σ	Reduced radii of curvature used in Hertzian contact calculation
Re	Reynolds number
S	Slide-roll ratio, defined as $2(u_1 - u_2)/(u_1 + u_2)$
T	Torque, various usages, with subscript indicating precise usage
T_{sp}	Torque lost to spin
U	Entrainment speed (average of u_1 and u_2)
V	Vehicle velocity, also Voltage in capacitance calculations
W	Normal load applied to contact
X	Reactance in capacitance calculations
Z	Number of roller elements in bearing-loss calculations, also impedance

Greek Letters

α	Contact angle between ball element and toroidal disc, also used to indicate pressure-viscosity coefficient
β	Contact angle between ball element and input conical disc
γ	Contact angle between ball element and output conical disc
ε_0	Permittivity of free space (electric constant) in capacitance calculations
ε_r	Relative permittivity of intermediary material in capacitance calculations
η	Transmission efficiency
λ	the angle of the axis of rotation relative to the input shaft
μ	Traction/Friction coefficient, also used to indicate mean in probability distribution
μ_c	Normalised Coulomb friction in LuGre tire friction model
μ_s	Normalised static friction in LuGre tire friction model
ν	Kinematic viscosity
v_s	Stribeck relative velocity in LuGre tire friction model
ρ	Density of oil
σ	Standard deviation in probability distribution calculations
σ_0	Rubber stiffness in LuGre tire friction model
σ_1	Rubber lumped damping in LuGre tire friction model
σ_2	Viscous relative damping in LuGre tire friction model
ψ	Dimensionless spin parameter
ω	Rotational velocity, various usages
ω_{sp}	Spin velocity
ω_{in}	Rotational velocity of input shaft
ω_{out}	Rotational velocity of output shaft
ω_{ball}	Rotational velocity of ball element shaft

VI. FIGURE LIST

VII. TABLE LIST

CHAPTER 1: INTRODUCTION

1.1 Background

1.1.1 Justification of Research

Transmission systems provide speed and torque conversions from a rotating power source to another device. They are crucial to the operation of a wide variety of mechanical systems in a vast number of different industries, including the automotive industry. Mechanically a transmission system in an automotive application is complex and is required to disconnect and connect the engine drive train from the wheels as required; reduce the rotational speed of the engine; and vary the transmission gear ratio as required by the driver to match the torque demanded at the wheels. Hence the efficient transmission of power from the engine to the wheels of an automotive vehicle is one of the greatest challenges facing automotive engineers (Dutta-Roy, 2004).

Transmission systems can be broadly divided into two categories: step-gear, in which the transmission has a discrete number of individual ratios; or continuously variable (CVT), where the transmission has an infinite range of ratios between two limits. Despite the inherent benefits of continuously variable transmissions, step-gear transmissions are traditionally more widely used in the majority of industries. One reason for this is that step-gear transmissions traditionally have higher transmission efficiencies. Furthermore in applications where the power source is flexible enough to provide a wide variation of operational torque and rotational speed, a continuously variable transmission would not provide a significant benefit.

Potentially a CVT could be used in a wide variety of applications. Perhaps the most technically demanding and potentially beneficial of these is the automotive industry. The majority of automotive vehicles manufactured today use fixed ratio transmissions, which results in the engine not operating at an optimum point at all times, reducing fuel efficiency and increasing emissions. Conversely, vehicles fitted with a CVT can offer improved performance and significant savings on fuel and engine emissions, just by allowing the vehicle's engine to operate within a higher efficiency envelope more frequently. Boos and Mozer (1997), whom

were some of the first researchers to methodically compare the benefits of CVTs, claim that a vehicle fitted with a CVT can achieve a reduction of more than one second when accelerating from 0-60mph, and a reduction in fuel consumption (and hence exhaust emissions) of at least 10%, whilst Arita (2000) claimed this figure could be as high as 10-20%. The recent global focus on reduced engine emissions indicates that there is a strong desire and need for this type of technology.

Developments in automotive technology, especially in internal combustion engine technology, have seen large improvements in fuel efficiency over the last 40 years. This is supported by Figure 1-1 (Schäfer, Heywood, and Weiss, 2006), which shows trends in the US automobile 'fleet' since 1970, and projections up to 2030.

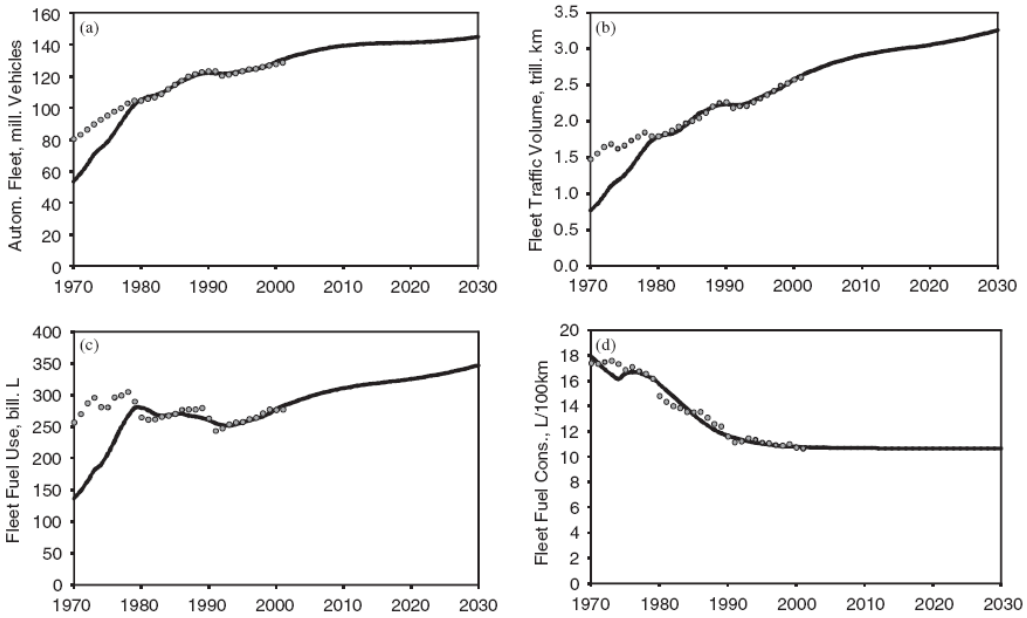


Figure 1-1: Trends of the US automobile fleet, historical data and projections

This Figure shows that despite large increases in fleet size (a) and distance travelled (b), there was actually a decrease in overall fuel usage from 1980 to the mid-1990s. This is entirely down to improvements in fuel efficiency, as proven by (d), which shows a decrease in fuel use per 100km travelled. More recently however, this trend is starting to slow, and since 2000, there have been virtually no further improvements in fuel efficiency. Transportation in general currently accounts for approximately a quarter of all emissions, and with global car ownership expected to double from 800million today to 1.6billion in 2030 (Ilett, 2004), this is a problem that needs to be addressed urgently.

One problem is that despite significant improvements in engine technology, transmission technology has remained largely the same for a number of years now, and only recently are improvements starting to be implemented. This is a major cause of inefficiency as traditional step-gear transmissions do not allow the engine to operate within an efficient regime for any significant length of time.

1.1.2 Overview of Current Continuously Variable Transmission Research

Continuously variable transmissions have been a very attractive alternative to automatic gearboxes for many years; however their potential has never been fully exploited and applied routinely into automobiles' drive-trains. Despite this, Continuously Variable Transmissions (CVTs) are generally viewed as theoretically perfect gearboxes. They provide a continuous, smooth variation of ratio, offering significant benefits for any application that requires a changeable transmission ratio, such as in automotive applications. In this application a continuously variable transmission can reduce engine noise, improve fuel efficiency, increase acceleration and improve ride smoothness. Early belt-type CVTs highlighted the potential of this technology but were never widely implemented due to material limitations and inherent design flaws that limited their torque capacity and durability (Kluger and Fussner, 1997).

The penetration of CVTs into the automobile industry has been slow. The main reasons have traditionally been considered to be excessive weight, limited torque capability in dynamic conditions, large inertia of rotating elements, limited durability and high cost in comparison to manual transmissions (Kluger and Fussner, 1997, and Heilich and Shube, 1983). In relatively recent years crucial research in material technology and design has seen significant improvements in CVTs, which have led to improved durability, efficiency and torque capacity.

The two main types of CVTs currently used in the transmission of automobiles (albeit sparsely) are belt/chain-type and toroidal traction drives. The former uses belts of various constructions which transmit the power from a set of conical pulleys located on the driving shaft to another set on the driven shaft. By changing the axial position of one or both pulleys, the radius of contact between the belt and the cones changes, varying the transmission ratio accordingly. The main problems involved with this type of CVT are the relatively low power transmitted, limited by the friction coefficient between belt and pulleys, the tensile strength of belt, and the wear of the elements due to the high dry or boundary friction forces present.

Toroidal traction drives transmit power by shearing a film of lubricant entrained between contacting elements. Current toroidal CVTs use either full or half toroidal input/output discs,

with toroidal or spherical intermediary elements. The main differences between different traction drive designs occur in the loading and synchronising mechanisms of the intermediary elements, whilst some designs also employ a planetary gear set to achieve infinite transmission ratios, and various systems for selecting neutral position and a reverse gear. Abrasive wear between contacting parts is eliminated as the metallic surfaces are separated by a fluid film; hence the durability is limited only by contact fatigue. Careful design of the device in order to limit the magnitude of the contact pressure, correct choice of the surface finish and hardness and the selection of the materials for the contacting bodies, can make the working life of these elements beyond that of the whole assembly. The fluid used in these applications is specially designed to have a high traction coefficient even at relatively large temperatures. To reduce the chance of contact failure, traction drives are designed to limit the shear stress applied to the lubricant. For example the traction coefficient of an industry-standard traction fluid is typically 0.1 at 40°C and 0.08 at 100°C, (Anghel, Glovnea, and Spikes, 2004) whilst traction drives are generally designed to not exceed a traction coefficient of 0.045 (Fuchs and Hasuda, 2004). This allows for overloading of the contacts or unexpected temperature rise, without the failure of the film.

The reported efficiency of toroidal CVTs is between 90 and 95% with a transmission efficiency of 85-90% (Newall at al., 2004 and Tanaka at al., 2004) depending on the particular construction. This efficiency is superior to an automatic gearbox, but lower than that of a manual gearbox (Lang 2004). The main advantage of this type of transmission thus relies on its ability to allow the engine to work at a higher efficiency, at any demanded load and speed, improving the overall fuel efficiency of the vehicle. A summary of the benefits discussed is shown in Table 1 (adapted from Harris, 2008).

Table 1: Summary of features and benefits of using a CVT in a automotive vehicle

Feature	*Benefit*
Step-less gear-change from rest to cruising speed	Eliminates “shift-shock” making the ride smoother
Keeps the engine in its optimum power range regardless of how fast the car is travelling	Improved fuel efficiency
Responds better to changing conditions, such as changes in throttle and speed	Eliminates gear hunting as a car decelerates, especially going up a hill
Less power loss in a CVT than a typical automatic transmission	Better acceleration
Can incorporate automated versions of mechanical clutches	Replaces inefficient torque converters

In summary traction-drive CVTs can offer a high power capacity, good durability and relatively compact design, making them ideal for the automotive industry (Kluger and Fussner, 1997). Traction drives are typically more expensive and less efficient than manual transmissions, and require a more complicated control mechanism than automatic transmissions. However a newer traction drive design, the Constant Power-CVT (CP-CVT) (Cretu and Glovnea, 2005), vastly simplifies the control operation by theoretically allowing a fully autonomous mechanical response to changing torque demands and road conditions. By improving the transmission efficiency of this design, and optimising it to meet technical demands, it has the potential to be a very strong alternative to current transmission systems.

1.1.3 The Constant-Power Continuously Variable Transmission Operating Principle

Recently a novel type of toroidal CVT has been developed that is capable of automatically adjusting the transmission ratio as a function of the resistive torque (Cretu and Glovnea, 2005). The device consists of two input discs, one conical, fixed to the shaft and the other toroidal, which has axial but not rotational mobility relative to the shaft. An inverted conical output disc is connected to the output shaft through a mechanism which is able to convert torque to axial force, such as a ball-screw. Between the input and output discs there are placed a convenient number (typically three-five) of spherical elements, which do not have a materialised axis of rotation. The arrangement of these parts is shown in Figure 1-2.

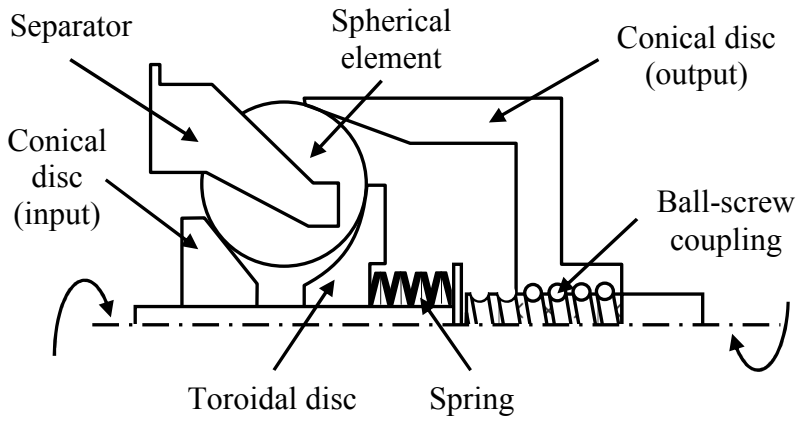


Figure 1-2: Design solution for a constant-power CVT

A schematic of the key components and forces, which reveals the functioning principle, is shown in Figure 1-3. A resistive torque (T) applied to the output shaft causes the output disc to move axially, forcing the balls to change their position relative to the input shaft, thus changing the position of the contact points on each ball element. This will cause a change of the

transmission ratio, the degree of which will vary depending on the geometry of the elements. The force produced by the ball screw coupling is balanced by a force applied to the toroidal disc (F_A), which also serves to load the contacts allowing traction to develop between elements. This force is provided by a carefully designed loading system, which in the simplest case can be linear with respect to toroidal disc displacement by use of a disc or coil spring. The balance of this force and the force produced by the ball-screw coupling means the CP-CVT is able to automatically adjust the transmission ratio to overcome any torque applied to the output, providing a potentially constant-power output.

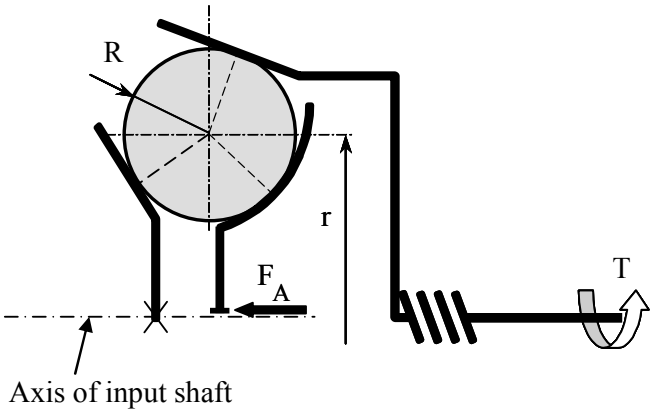


Figure 1-3: Functioning principle of CVT

A possible design arrangement of the CP-CVT is shown in Figure 1-4. This figure highlights several benefits of this design, such as co-aligned input and output shafts, relative compactness, and the overall simplicity of the device. A ball separator or cage (shown) is also required to ensure the spherical elements rotate about their own axis and do not contact one another. The separator shown in the figure is perhaps the simplest possible design; however such a design does potentially yield a high source of inefficiency, and ideally should be improved.

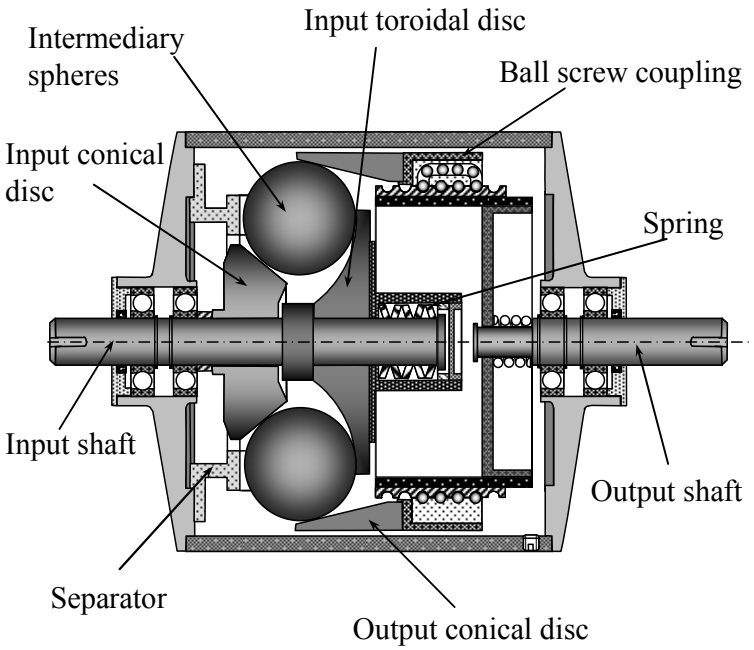


Figure 1-4: Intended CP-CVT layout

Perhaps the biggest criticisms of current traction drive CVTs are the high tolerance and strength requirements of the primary components, which leads to increased manufacturing costs. In addition to this they typically require a complex control system to adjust the intermediate components in order to change the transmission ratio. However the CP-CVT design shown overcomes several of these inherent flaws. By altering the transmission ratio automatically based on the torque load applied to the output, the CP-CVT can be potentially fully automatic and hence require no external hydraulic or servo control mechanism, reducing costs and simplifying the overall system. The axes of rotation of the intermediary elements are not materialised, which means that there is no need for a complex and costly synchronisation mechanism. This design offers a relatively compact solution when the load behind the toroidal disc is generated by a spring and a wide, continuous range of transmission ratio, making it a very attractive alternative to current transmission systems.

1.2 Purpose of Research

Although the CP-CVT design has existed in some form for almost three decades, the design has not yet been optimised, and hence is far from ready to be implemented in a real automotive application. Whilst some attempts have been made to optimise the design and each of the component dimensions in isolation (Cretu and Glovnea, 2006), no attempt has yet been made to take the design from the concept stage to a usable product, which is one of the primary purposes of this thesis. In order to achieve this, a number of new and existing design optimisation techniques are used to analyse the CVT from a design and tribological prospective. These techniques look at optimising each dimension and design in parallel in order to meet multiple objectives, culminating in a design that is essentially ready to be manufactured and implemented.

The validity of the optimised design is demonstrated through the use of a "ground-up" simulation that attempts to model the behaviour of the CVT in a real automotive application using multiple fundamental theories and models including tire friction and traction behaviour.

Additional complementary research looks at the accuracy of the tire friction models through the use of a specially design tire friction test rig. Furthermore, a monitoring system is proposed for this particular CVT design (and similar) that is capable of continuously checking the contact film thickness between adjacent elements to ensure that there is sufficient lubrication to avoid metal-on-metal contact. The system, which is based around electrical capacitance, requires the knowledge of the behaviour of the electrical permittivity at increased pressure. This behaviour is studied through the use of an experimental test rig.

1.3 Layout of Thesis

This thesis consists of 10 main chapters, which are summarised here:

Chapter 1: Introduction: *This chapter introduces the main research subject and discusses why it is considered necessary.*

Chapter 2: Literature Review: *This chapter looks at existing research conducted around the main subject area and associated research subjects that are used in this thesis.*

Chapter 3: Initial Design Process: Quality Function Deployment: *This chapter discusses the first stage of the design process, looking at the CP-CVT from customer-orientated-design prospective.*

Chapter 4: Efficiency Optimisation: *This chapter optimises one particular technical requirement of the CP-CVT, developing equations to predict the transmission efficiency. A novel optimisation algorithm is also developed to dimensionally optimise the CP-CVT.*

Chapter 5: Multi-Criteria Dimensional Optimisation: *This chapter uses two separate algorithms to simultaneously optimise the CP-CVT based on multiple technical requirements, which were determined during the initial design process.*

Chapter 6: Vehicular Simulation: *This chapter shows the development of a simulation tool that allows the CP-CVT to be modelled in vehicular environment. The previously optimised dimensions are incorporated to determine if the CP-CVT can realistically operate as desired.*

Chapter 7: Complementary Research: Rubber friction: *During the development of the simulation shown in chapter 6, it was found that current tire friction models maybe representative, but generally have little relation to physical, measurable parameters. This stand-alone chapter uses a specially designed rubber-friction test rig to analyse the effect of microscopic and macroscopic asperities on friction and how these effects might be incorporated into existing tire friction models.*

Chapter 8: Complementary Research: Proposed Method for Contact Film Thickness Measurement: *This chapter proposes a method for monitoring contact conditions within the CP-CVT during operation. This method could potentially give an advanced warning of contact failure, helping to prevent surface damage of the contacting components, which are potentially the most expensive parts of the CP-CVT*

Chapter 9: Design Critique and Prototype Design: *This chapter presents a design critique of the CP-CVT, discussing problems that have been solved through the use of the design theories presented, whilst definitive design solutions are proposed for problems that still exist. Additionally the design of a proposed prototype test-rig is shown.*

Chapter 10: Conclusion: *The final chapter contains a summary of all the findings and conclusions of this thesis, together with suggestions for future work on the subjects.*

CHAPTER 2: LITERATURE REVIEW

2.1 Transmission Technology

2.1.1 Overview of Transmission Technology

A transmission mechanism is required in any application that requires a conversion in the speed or torque from a rotating power source to another device. The majority of power producing machines, such as turbines, internal combustion engines and electric motors produce power in the form of rotational motion. The operating characteristics of these power sources vary considerably depending on their type and size, and often require a transmission of some sort in order to transform the speed or torque into a usable form (Childs, 2004). They are hence crucial to the operation of a wide variety of mechanical systems in a vast number of different industries.

As mentioned earlier, transmission systems can be divided into two categories: discrete-ratio, or continuously-variable. The simplicity of discrete-ratio transmissions means they are far more widely used in the majority of industries, for which a limited or fixed range of ratios is sufficient. This typically occurs in applications where only a narrow range of torque and speed variation is required or where the power source is sufficiently flexible already. An example of this is on a standard push bike, in which the power source is a human body. The human body is remarkably flexible, as proven by the fact that even with a single-geared bike the body is capable of moving the bike from rest to high speeds. Hence, whilst there have been attempts to apply CVT technology to a push bike, such as the Nu-Vinci CVT (Hogan and Donahue, 2008), they have never been widely accepted, simply because they are not required. Historically, manual or fixed ratio gearboxes also offer a much higher transmission efficiency (94-97%, Lang, 2004), in comparison to CVTs (75-95%, Kluger and Long, 1999), however recent technological advances have closed this gap considerably.

One particular industry that would benefit greatly from the use of a continuously variable transmission is the automotive industry. Improvements in automotive technology, especially in internal combustion engine technology, have seen a large reduction in the fuel used per distance travelled (Schäfer, Heywood, and Weiss, 2006). In recent years however, this trend is starting

to slow as engine efficiency improvements have reached a plateau. Indeed some sources actually claim that the opposite is true in recent years, with the average MPG of cars purchased more recently actually decreasing, as shown in Figure 2-1 (Greene and DeCicco, 2000).

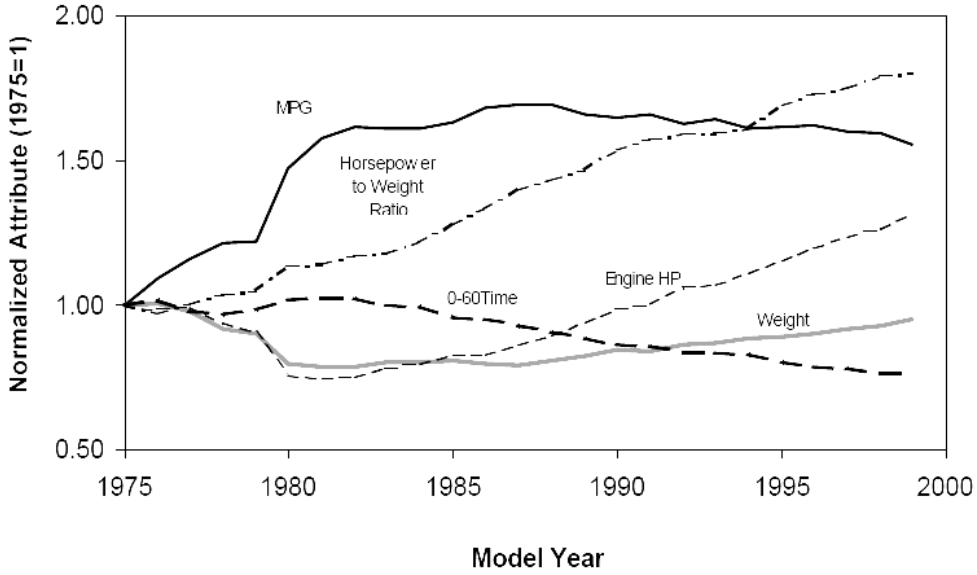


Figure 2-1: Trends in fuel efficiency, engine power and vehicle weight from 1975-1996

Looking at this figure, the reason for the recent decrease in average vehicle MPG is simply because of the increases in vehicle power (demanded by the customer) and weight (required for increased safety). Furthermore since 1995 there has been a surge in the number of larger vehicles sold such as SUVs (Hellman and Heavenrich, 2001), which have a significantly higher mass, and hence use more fuel per distance travelled.

All of these factors will ultimately lead to a major problem, especially as fuel usage is projected to continue to increase in the near future (Schäfer, Heywood, and Weiss, 2006). Dwindling fuel reserves and increases in population and car ownership imply that this is a problem that needs to be addressed urgently. Ignoring the use of electric motors, which only transfer emissions from one source to another, and the use of hydrogen engines, for which an infrastructure is not in place, the best way fuel efficiency can be currently improved further is by increasing the amount of time an engine remains in its peak efficiency regime. The importance of this is highlighted in Figure 2-2 (adapted from Bucknell, 2006).

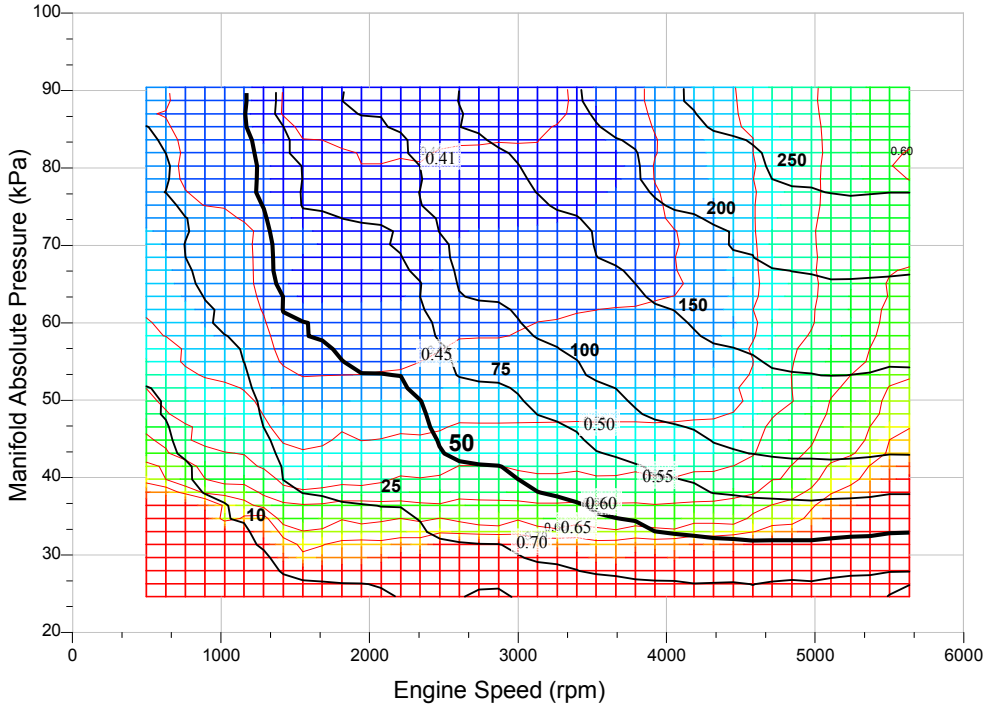


Figure 2-2: Effect of engine speed and manifold pressure on power and fuel consumption

This rather complex figure (taken from a real engine) simply shows the effect of manifold absolute pressure (MAP) and engine speed on the specific fuel consumption (red iso-contours, measured in lbs of fuel per bhp-hrs) and engine power (black iso-contours, measured in bhp). The specific fuel consumption simply gives an indication of how efficiently the engine is running, with lower values indicating a better fuel efficiency. From this figure it can be seen that for any particular power output, there remains a vast number of different operating points the engine could run at. For example, following the 50bhp power iso-contour (highlighted in the figure) the engine could either run at a high engine speed, with a lower MAP, giving a specific fuel consumption of >0.7lb/hp-hr, or alternatively at a low engine speed and higher MAP, giving a specific fuel consumption of approximately 0.43lb/hp-hr. Thus even for the same power output, there is a difference in fuel consumption of almost 40% depending only on the operating regime of the engine. This is potentially a major cause of vehicular inefficiency for traditional step-gear transmissions, which are physically incapable of allowing the engine to operate within an efficient regime for any significant length of time, something which a CVT is capable of achieving. Although a difference of 40% is the hypothetical maximum difference between inefficient and efficient operation, genuine figures based on experimental analyse claim that the use of CVT can reduce fuel consumption in a typical automobile by anywhere from 10%-20% (Boos and Mozer, 1997; Arita, 2000).

2.1.2 Manual and Automatic Transmissions

Manual transmissions, which are historically more popular in Europe, require that the correct gear ratio is manually chosen by the driver, increasing the amount of effort the driver has to exert, but also increasing perceived control. The earliest manual transmission design could be considered the "sliding mesh gearbox", in which gears are engaged by the sliding of the gears on a splined shaft using a selector fork. This has the disadvantage of causing the gears to 'grind' (gear-clash), which occurs when the associated shafts are rotating with different speeds. Modern manual transmissions overcome this through the use of "constant mesh gearboxes", in which the gear pairs are always interlocked. Each gear pair is selected though a dog-mesh (a type of clutch, Childs, 2004), whilst a synchroniser can also be employed to ensure the shafts are rotating at similar speeds when the clutch engages. By ensuring the gears are always engaged, these types of gear boxes have the significant advantage of reducing gear-wear, which can happen when gear-clash occurs.

Automatic transmissions conversely require no input from the driver to select an appropriate gear using advanced algorithms to change gear automatically based upon throttle position, engine load and vehicle speed. More complex in design, automatic transmissions typically utilise epicyclic (planetary) gear systems to provide various ratios using electro/hydraulic actuators. A torque convertor is also typically employed to permit a degree of 'slip' allowing the vehicle to move away from rest and to dampen engine excitation. The torque convertor itself is a large source of inefficiency within the drive train, and is one of the reasons both fuel efficiency and performance are inferior on vehicles fitted with automatic transmissions.

Despite the extra effort required, nearly 80% of vehicles sold in Europe come equipped with a manual transmission (Bharat Book Bureau, 2006). It is thought that the reason for this is that motorists with an interest in performance prefer to select the correct gear themselves (control), whilst those whom are more concerned with fuel economy believe that automatic transmissions deliver inferior economy (Bharat Book Bureau, 2006). The latter reason is generally true, with properly operated manual transmission vehicles offering an improvement of about 5% to 15% over automatic transmissions depending on driving conditions and style of driving (Kluger and Long, 1999). Outside of Europe this trend is not repeated, especially in Northern America, where drivers almost universally prefer automatic transmissions. Because of this, it can be reasonably expected that CVT technology is more likely to be adopted in North America before Europe.

2.1.3 Continuously Variable Transmissions

The concept of Continuously Variable Transmission has existed for over a century, with the first European patent being taken out in 1886. Even before this, Leonardo Da Vinci had his own idea about the concept as early 1490 (Harris, 2008), shown in Figure 2-3

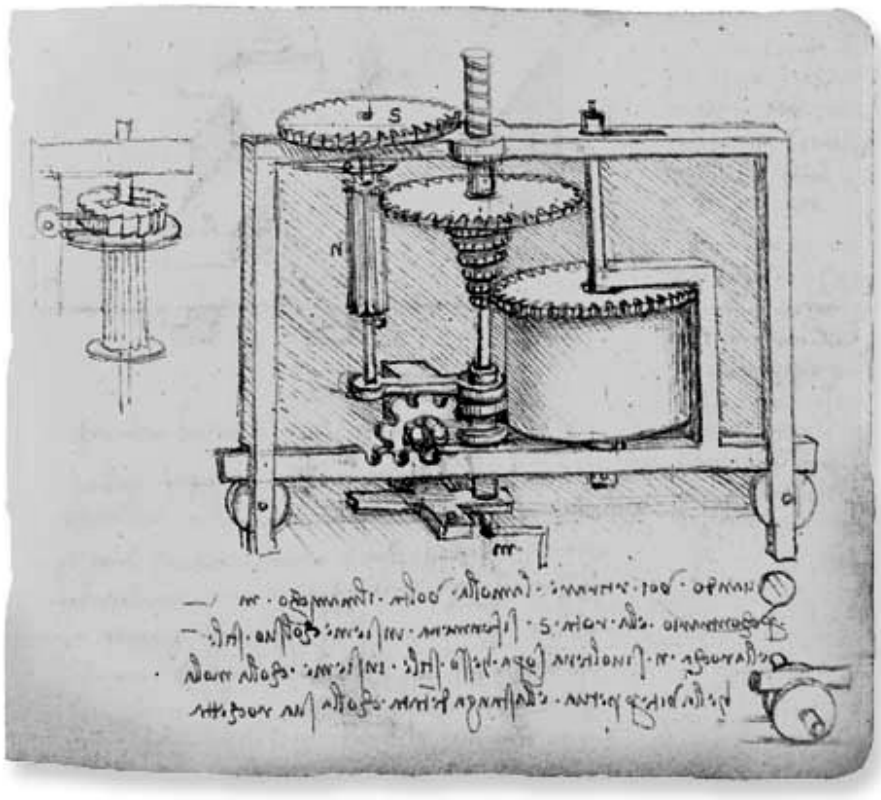


Figure 2-3: One of the earlier designs for a Continuously Variable Transmission

A good history of the development of traction drives is presented by Heilich and Shube (1983). It is claimed here that CVTs have enjoyed moderately wide-spread industrial use since the 1930s, whilst the automotive industry, notorious for being slower to react to new technologies, never really adopted their use, despite the obvious advantages discussed previously. The reluctance of the automotive industry to react to new technologies is highlighted in Figure 2-4 (from Hellman and Heavenrich, 2001).

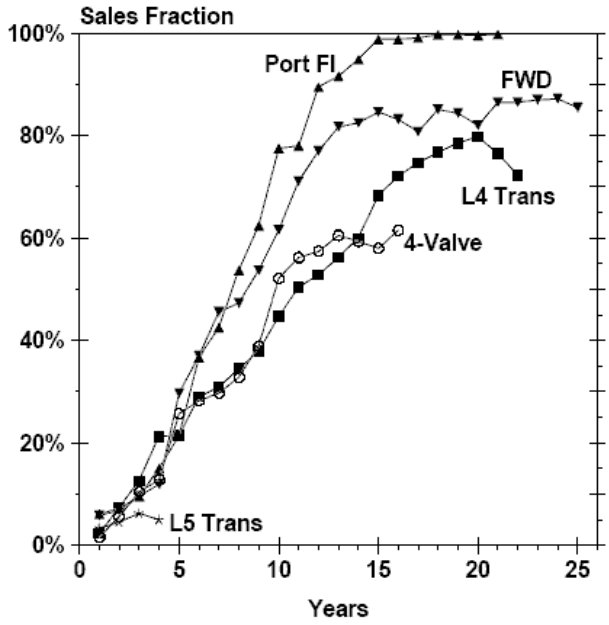


Figure 2-4: Market penetration of newly developed automotive technology

Several of the technologies shown in this figure such as 4-values per cylinder, fuel injection and front-wheel drive (FWD) are now taken for granted largely due to their benefits to consumers. Despite this, it still took the industry a long time to adopt them after they were initially developed. Similarly, a friction drive for automotive use was developed and implemented as early as 1900, marketed to consumers as "The Friction Drive Car" (Heilich and Shube, 1983), however even now the technology remains a rarity within the market meaning it possibly holds the record for the longest period from initial concept to wide-spread market penetration. The reasons for this are numerous and include material limitations and limited torque capacity (Kluger and Fussner, 1997), and production cost and durability (Heilich and Shube, 1983). Only recently have these problems been somewhat overcome through the use of traction fluids, and advancements in material technology. The next hurdle is thus convincing drivers of the advantages of using CVTs, whom (it is claimed) partially reject their use because of the *perceived* lack of acceleration and control in comparison with standard transmission systems (Wicke et al., 2002).

Although Heilich and Shube present a good historical account of traction drives, a better overview of the current level of CVT technology is presented by Lang (2000). It should be noted however that the significant rate of advancement of this technology in recent years is highlighted by the omission of certain recent designs from this paper, written in 2000. Within this paper is a comparison of the efficiencies of various CVT types together (Table 2), in addition to the typical efficiencies of the more widely used step-gear, automatic transmission (Table 3) (both Lang, 2000). It should be highlighted that this is only the mechanical

transmission efficiency and furthermore certain CVT designs have improved on these figures since they were published.

Table 2: Typical CVT efficiencies

CVT Mechanism	*Efficiency Range*
Rubber Belts	90-95%
Steel Belts	90-97%
Toroidal Traction	70-94%
Variable Geometry	85-93%

Table 3: Typical automatic transmission efficiency by gear

Gear	*Automatic Transmission Efficiency*
1st	60-80%
2nd	60-90%
3rd	85-95%
4th	90-95%
5th	85-94%

Although these figures are relatively low in comparison to a standard manual transmission (approximate mechanical efficiency of 97% (Lang, 2000), it should be noted that when coupled with an internal combustion engine, the CVT's ability to allow the engine to operate more efficiently provides notable savings on fuel, as found by both Boos and Mozer (1997) and Arita (2000).

2.1.4 Traction Drives

Although steel-belt designs are slowly becoming an option on certain automobiles, and despite the historic relative inefficiency of toroidal traction drives (highlighted in Table 2), many researchers still believe that the transmission system that shows the most potential is the toroidal or half-toroidal traction drive. The advantage of traction CVTs over steel-belt designs is the lack of contacting parts, which are separated by a traction fluid, significantly increasing the life of the transmission. Despite this, many early attempts that were made to apply the full toroidal CVT to automotive power-trains failed mainly due to durability and controllability problems (Cahn-Speyer, 1957 and Carson, 1975). Whilst the durability issues have been solved through advances in technology since these conclusions were made, the issue of controllability still

affects a number of traction drive designs. This controllability issue can either be due to issues with transmission ratio control (discussed later), or actually controlling and synchronising the working parts of the device itself.

Traction drive devices usually consist of a number of toroidal or half-toroidal input and output discs, which allow intermediary elements to form a variable radius across a toroidal surface, thus altering the transmission ratio. The intermediary element is generally either spherical (Cretu and Glovnea, 2005; Keun and Park, 2004; and Pohl et al., 2004) or roller-shaped (Tenberge and Mockel, 2002; Fuchs, Hasuda and James, 2002; Akehurst et al., 2001; Brockband and Heumann, 1999). Perhaps the most advanced designs in terms of research development and investment are Torotrak's IVT, the Milner CVT and Nissan's half-toroidal traction CVT.

2.1.4.1 Torotrak Full-Toroidal Traction Drive

Torotrak have released over 40 technical papers regarding their CVT design and associated technologies, and hence there is an abundance of literature available on the design. The device consists of 2 sets of toroidal input and output discs with roller elements, coupled with an epicyclic gear set as shown in Figure 2-5 (Fuchs, Hasuda and James, 2002).

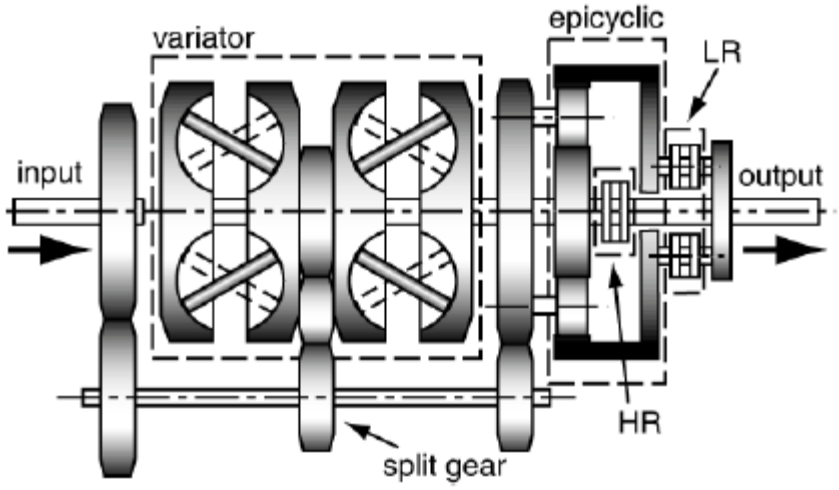


Figure 2-5: Torotrak Full Toroidal CVT

The epicyclic gear set allows the transmission to operate in both directions and with a zero net speed output (geared neutral). Further details regarding the concept of this geared-neutral are discussed in (Brockband and Heumann, 1999), which also serves as one of the earliest papers to introduce the design. The transmission ratio is controlled by altering the positions of the roller

elements, which adjust their angle automatically due to side-slip within the elasto-hydrodynamic (EHD) contact. In addition to this there is a pair of clutches that allow the transmission to operate in either high or low regime, i.e. high or low transmission ratio (Fuchs, Hasuda and James, 2002).

In addition to discussing the concept of geared neutral Brockband and Heumann, (1999) also discuss some of the aspects of control of the Torotrak-Infinitely Variable Transmission (T-IVT). They state that the T-IVT is a "torque controlled transmission with the driver demanded wheel torque achieved by controlling the pressure on each side of the hydraulic pistons connected to the rollers in the Variator." In order to achieve this, they reveal that extensive modelling was required to "provide the detailed relationship between wheel torque demand and the Variator reaction torque and reaction pressure." The complexity of this control mechanism highlights one of the perceived deficiencies of the design. Furthermore, it is critical to the operation of this design that all intermediary rollers operate in unison to ensure that one is not driven by another. Control of the roller's tilting angle is one of the key issues in the development of this type of CVT (Zhang, Zhang and Tobler, 2000). It is perhaps the difficulty of this and the overall complexity of the design that has prevented the T-IVT from entering mainstream production, despite significant investment and research. The sheer number of papers regularly written regarding the control of the T-IVT indicates that this issue has still not been fully resolved (Zhang and Dutta-Roy, 2004; Field and Burke, 2005; Greenwood, 2007; Fuchs, Hasuda and James, 2002).

More recently, research on the Torotrak design has focused on analysing and assessing the fuel economy benefits of vehicles fitted with the T-IVT (Burtt, 2007; Brockbank, 2009; Brockbank and Greenwood, 2009; Dick, 2010) whilst further research includes discussions about drivability (Wicke et al., 2002), the prediction of contact losses (Newall and Lee, 2003), and overall efficiency modelling (Newall et al., 2004), which are discussed later.

2.1.4.2 Milner CVT

Another alternative traction-drive design is the patented Milner-CVT (M-CVT), introduced properly by Akehurst et al. (2001). Like the CP-CVT, the M-CVT design uses three or more spherical intermediary rollers to transfer drive from an inner raceway to a carrier assembly. The transmission ratio is changed by altering the axial separation of outer race set. Again, like the CP-CVT, this can be designed to have co-aligned input and output shafts, as shown in Figure 2-6.

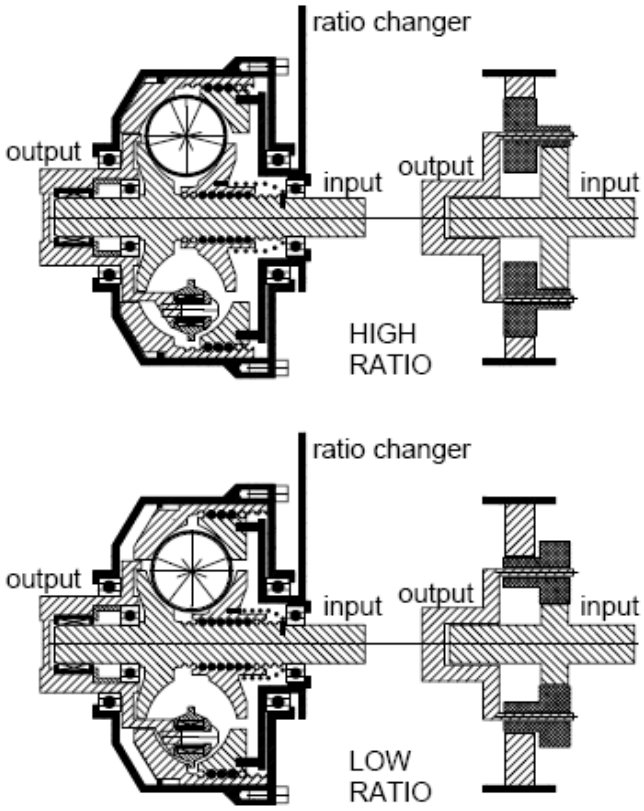


Figure 2-6: Schematic representation of Milner-CVT

Further research on the Milner-CVT is presented in Akehurst, Parker and Schaaf (2007), where a basic dynamic model is presented, and the kinematics of the design are discussed in more detail. This model is inelastic and implemented in SIMULINK, which allows the position, speeds and forces of each component to be calculated. The majority of the work presented in fairly low-level and is designed more to disseminate the design to the wider research community.

An optimisation approach to the Milner-CVT is shown in Akehurst et al. (2009). This conference paper again serves to introduce the operating principle of the design, and expands on this to develop a tool, based on a multi-objective genetic algorithm. The results presented are basic at best, which perhaps indicates that the design optimisation process used encountered some difficulties. Furthermore, little information is actually given about the nature of the evaluation functions used, how they are calculated and how they are implemented. The graphs shown indicate that the predicted efficiency of this design varies widely from approximately 70-95%, depending on the transmission ratio, and scale (dimensions).

This particular design could be considered the closest competition to the CP-CVT, given the similarities between the designs, however significant differences in the nature of the contacting

elements, and the control of the transmission ratio, make the CP-CVT an overall more adaptable design. Several novel ideas from the M-CVT highlight useful additions that can be made to a number of other CVT layouts. These include:

- The ability to be made fully reversible with respect to both torque and rotation directions
- The possibilities of including a built-in freewheel for uni-directional applications.
- The inclusion of a "passive, direct-drive top gear lock-up", which it is claimed can increase top-gear efficiency by about 10% (Akehurst et al., 2001)

2.1.4.3 Other Traction Drive Designs and Competitors

Although other traction drive and CVT designs might appear to be the greatest threat to successful development of the CP-CVT, other emerging technology could also eliminate the need for CVTs altogether. This includes the use of electric-type engines, for which the benefits of a CVT are substantially reduced, and more recently, the use of DCT (dual-clutch technology). Dual clutch technology perhaps poses the greatest risk to wide-spread CVT acceptance since it is more similar to existing vehicle control, which drivers have grown accustomed to and are reluctant to abandon. The advantage of DCT technology over automatic transmission is that the inefficient torque convertor is discarded in favour of automated, hydraulically-controlled clutches, improving fuel efficiency and performance. The basic principle uses two clutches that alternate usage between gear changes, ensuring that there is always a connection from the engine to the driven wheels. Although this type of transmission was originally invented several decades ago, it is only recently becoming widely implemented, with some optimistic predictions stating that by 2014, 15%, or even 20% of light vehicles manufactured in Europe could be fitted with a DCT (DCTfacts.com, 2008). Certain established CVT researchers have already started switching focus towards DCT (Zhang et al., 2005). The advantage of this technology is that it returns a degree of control to the driver, operating in either full-automatic or semi-automatic regimes, whilst still improving efficiency over automatic transmissions. Given that DCTs still employ fixed, discrete ratios, they still suffer the same problems discussed previously. The challenge for CVT designers is hence not necessarily in further development of the technology, but in convincing drivers of the benefits.

2.1.4.4 Nerac Report

A Nerac[1] report conducted for the purposes of exploring the intellectual property potential of the CP-CVT revealed the closest perceived rivals in terms of design and patent. Aside from certain belt-driven infinitely variable transmissions that employ a ball screw for automated control, the greatest threat to the CP-CVT was considered to be spherical element traction drives, such as the Milner-CVT described earlier. Additional patent threats were considered to be the Nissan/NSK CVT shown in Figure 2-7 (Tenberge and Mockel, 2002), and the Torotrak CVT, discussed previously, both of which are considered fundamentally different enough in design and operation that patenting is still a distinct possibility.

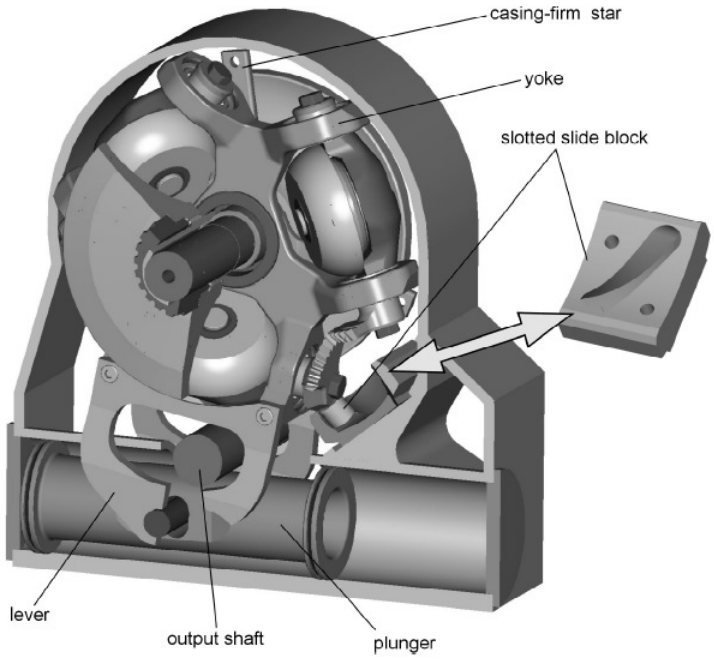


Figure 2-7: Nissan/NSK developed Toroidal CVT with compact roller suspension

2.1.5 CVT Ratio Control Strategies

With the use of a CVT, a car's engine speed and operating point can be controlled independently of vehicular speed. Determining the optimum operating point for static situations is a relatively simple task, determining the optimum control strategy for transient conditions is far more difficult. As mentioned previously, this issue of control is a major obstacle that needs to be overcome for any CVT design to be successful. Because of this, there has been a

[1] Nerac is a Research and Advisory Firm that delivers custom research and analysis for companies developing innovative products and technologies

significant amount of published work that attempts to offer solutions or strategies in relation to CVT ratio control. The most common approach is to offer a speed-envelope to the driver, who controls the specific operating point through use of the accelerator (Figure 2-8).

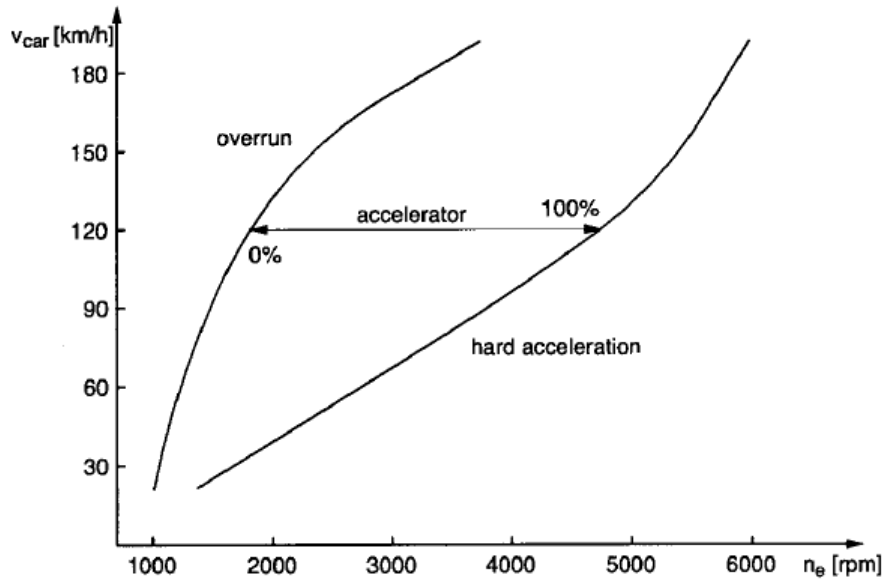


Figure 2-8: Speed envelope approach to ratio control (Pfiffner and Guzella, 2001)

The driver is given the acceleration curve he/she desires within a predetermined operating envelope. The lower line (hard acceleration) affects the driveability of the vehicle, whilst the upper line indicates the desired engine speed when cruising or overrunning ('engine-braking'). An alternative approach is to have a single operating curve (Pfiffner and Guzella, 2001). This solution requires some compromise between efficiency and driveability. The driver now indicates the desired power through the accelerator pedal, and the system adjusts the ratio accordingly to operate on a predetermined curve. This interpretation of pedal significance is supported by (Bucknell, 2006) who claims pedal position is a perfect indicator of desired acceleration/force. Whilst this solution could potentially offer a greater fuel economy by essentially restricting the available power at any moment, the driveability of the vehicle suffers. An extension of this is to offer two (or more) operating curves that the driver can manually select between to provide either increased performance, or improved fuel economy (Pfiffner and Guzella, 2001). One of the perceived deficiencies in vehicles fitted with a CVT is the lack of *apparent* acceleration in comparison with standard transmission systems (Wicke et al., 2002). This is partly due to the lack of an expected increase in engine noise that normally accompanies acceleration in step-gear transmissions. Whilst attempts have been made to simulate this effect, it invariably comes at a cost of increased fuel usage.

Another potential problem with ratio control strategies is the difficulty in characterising 'driveability'. When attempting to produce an operating envelope or curve, it is relatively easy to determine an optimum efficiency function, however determining an optimum driveability function is ambiguous at best. Whilst attempts have been made to characterise the driveability during transient conditions with varying degrees of success (Wicke, Brace and Vaughan, 2000; and Smith et al., 2004), improving drivability usually consists of assigning an 'arbitrary function…with a higher engine speed for all throttle inputs' (Smith et al., 2004). One attempt to quantify the entirely subjective concept of 'driveability' involves measuring drivers' perceived acceleration, with some promising results, including the use of a 'jerk value', which is the derivative of acceleration with respect to time (Wicke, Brace and Vaughan, 2000). However in simple terms, a vehicle could be described as having perfect drivability if it responded precisely and immediately to the driver's instructions (within the limitations of the vehicle). Even without a complex control mechanism the CP-CVT should still be able accomplish this. Quick, sharp increases in demanded torque result in the engine speed increasing as the CP-CVT adapts to find a new operating point, whilst for slower, steadier increases in demand the CP-CVT will constantly react to maintain a constant engine speed up to the transmission ratio limit. Hence the balance between engine efficiency and performance is controlled entirely by the driver's use of the pedal.

2.2 The Constant Power CVT

2.2.1 Overview

The operating principle and aspects of the Constant Power CVT (CP-CVT) have been presented in a number of technical papers (Cretu and Glovnea, 2005 and 2006, Glovnea and Cretu, 2006). Given the unique and specific nature of the design, all of the available literature has been written by the same authors.

The earliest paper that described the operation of the CP-CVT in detail was presented by Cretu and Glovnea (2005). This paper begins by giving a brief overview of the improvements that can be made to a vehicle's performance through the use of a CVT, then moves onto discuss the general operating principle of traction drives that utilise the elastohydrodynamic (EHD) concept. This discussion is only brief, and is examined in far more detail in tribology-specific papers. The majority of the paper thus serves as an introduction to this new design, which consists of two input discs, one conical and the other half-toroidal, an inverted conical output disc, and a number of spherical intermediary elements. The paper is keen to highlight the advantages of this design, which are claimed to be the elimination of the need to control the axis of rotation of the balls elements, and the constant-power aspect of the device.

Further work on the CP-CVT was presented in Cretu and Glovnea (2006), which looks at geometrically optimising the device. In addition to a smaller introduction on the concept of the design, the paper goes on to discuss the effect of changing the component dimensions. The only specific properties that are analysed include the transmission ratio, torque output and power output, ignoring critical factors such as efficiency or mass. Furthermore, each property is only analysed in isolation looking at the influence of one or two dimensions. This, as shown in later sections, is not a valid approach to optimisation, and because of this certain conclusions made in this paper regarding ideal component dimensions have later been found to be invalid.

Additional discussion of the CP-CVT design is shown in Glovnea and Cretu (2006). Although large parts of this paper are similar to what has been discussed previously (operating principle and geometric optimisation), a new and novel concept is introduced in the form of a double caged variation of the basic design. This concept aims at using multiple sets of spherical intermediary elements to increase the number of contacting points (and hence the power capacity), without significantly increasing the overall size. As with the earlier geometrical optimisation paper, a number of suggestions are made regarding the component dimensions, which are largely irrelevant now. The efficiency of the CP-CVT is also discussed in reference to a criticism of the design, specifically the high losses that occur due to the ball separator (cage). The paper attempts to respond to criticism, stating that "in the case of automobile

transmissions, while high efficiency is always desirable, it should not be the critical factor in determining whether a CVT should be used". Although this maybe partially true, the fact that these losses are high enough to warrant discussion implies that they are of concern, and need to be addressed.

2.2.2 Kinematic Analysis of the CP-CVT

In Cretu and Glovnea (2005), in addition to introducing the concept of the CP-CVT, the authors also present a mathematical analysis of the kinematics of the device. Although this paper was the inspiration for equations discussed herein, each of them has been derived from first principles based on what is required for this research; and hence there is some discrepancy with the final equations shown here. The equations derived based around five fixed input dimensions that can fully describe the key components of the CVT, two angular (β and γ) and three radial (R, R_1, r_0), in addition to the variable angle α, which controls the immediate properties of the transmission, as shown in Figure 2-9.

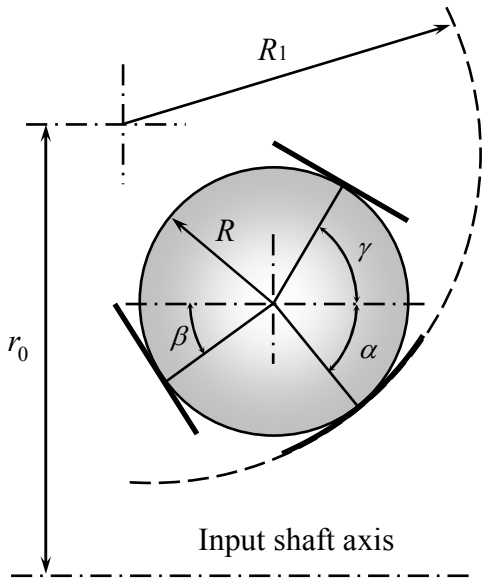


Figure 2-9: Explanation of dimensions of CP-CVT

These five dimensions describe and dictate the majority of the properties of the CP-CVT, such as its mass, transmission ratio, efficiency and torque output.

2.2.2.1 The Instantaneous Tangential Velocities of the Input Contact Points

The tangential velocities of the contact points between the ball elements and the toroidal and conical input discs are simply a function of their distances from the centre of the input shaft, and

the rotational speed of the input shaft (ω_{in}). These distances will change depending on the instantaneous position of the ball, which can be described by the distance from its centre to the input shaft (r). This distance can be shown to be a function of the contact angle between the ball and the input toroidal disc (α), as shown in Figure 2-10.

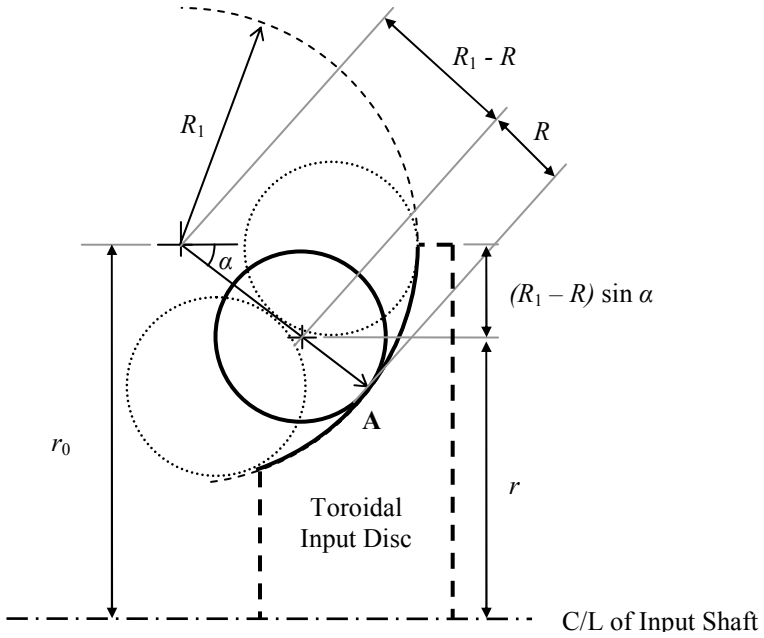


Figure 2-10: Relationship Between α and r

Hence:

$$r = r_0 - (R_1 - R)\sin\alpha \tag{2.1}$$

Knowing this, the tangential velocities of the discs ($u_{in\text{-}A}$ and $u_{in\text{-}B}$) at the input contact points **A** and **B** can be calculated as shown in Figure 2-11. Hence:

$$u_{in-A} = \omega_{in}(r - R\sin\alpha) \tag{2.2}$$

$$u_{in-B} = \omega_{in}(r - R\sin\beta) \tag{2.3}$$

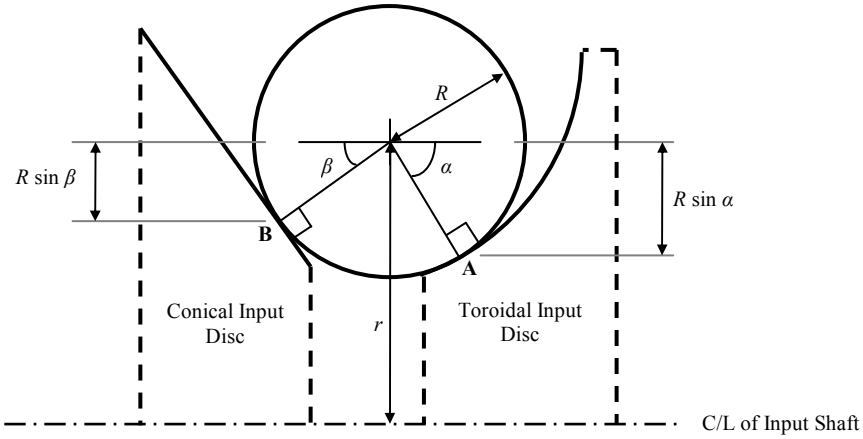


Figure 2-11: Calculation of Input Tangential Velocities (ω_1)

2.2.2.2 The Angle of the Axis of Rotation of the Ball Elements

As the position of the ball element changes, so too will the position of the contact points **A** and **B**. Since the shapes of the two input discs are different, the contact points will not always be inline, hence causing the ball element to rotate around an axis that is not always parallel to the input shaft. This is after all the principle behind the changing ratio between the input and output rotation speeds.

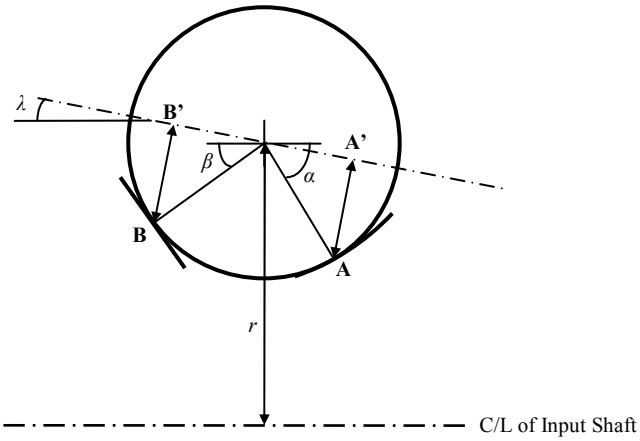


Figure 2-12: Calculation of Input Tangential Velocities (ω_2)

If the rotational speed of the ball element (about its own axis of rotation) is ω_{ball}, then the tangential speeds of the contact points **A** and **B,** (based on the rotation of the ball) are:

$$u_{ball-A} = \omega_{ball}[R\sin(\alpha - \lambda)] \quad (2.4)$$

$$u_{ball-B} = \omega_{ball}[R\sin(\beta + \lambda)] \quad (2.5)$$

If it is assumed that there will be negligible slip between the ball element and the discs, then the tangential speeds at the surfaces of the ball element and input discs can be considered to be approximately equal. Knowing this, the angle of the axis of rotation λ can be calculated by combining the Equations 2.2 to 2.5

$$\frac{(r - R\sin\alpha)}{[R\sin(\alpha - \lambda)]} = \frac{(r - R\sin\beta)}{[R\sin(\beta + \lambda)]} \tag{2.6}$$

Eliminating R and expanding using compound angle identities:

$$\frac{(r - R\sin\alpha)}{\sin\alpha\cos\lambda - \cos\alpha\sin\lambda} = \frac{(r - R\sin\beta)}{\sin\beta\cos\lambda + \cos\beta\sin\lambda}$$

Dividing by ' $\cos\lambda$ ':

$$\frac{(r - R\sin\alpha)}{\sin\alpha - \cos\alpha\tan\lambda} = \frac{(r - R\sin\beta)}{\sin\beta + \cos\beta\tan\lambda}$$

Cross-multiplying and expanding:

$$r\sin\beta + r\cos\beta\tan\lambda - R\sin\alpha\sin\beta - R\sin\alpha\cos\beta\tan\lambda$$
$$= r\sin\alpha - r\cos\alpha\tan\lambda - R\sin\alpha\sin\beta + R\sin\beta\cos\alpha\tan\lambda$$

Simplifying:

$$\tan\lambda[r(\cos\beta + \cos\alpha) - R\sin(\alpha + \beta)] = r(\sin\alpha - \sin\beta)$$

Hence:

$$\lambda = \tan^{-1}\frac{r(\sin\alpha - \sin\beta)}{[r(\cos\beta + \cos\alpha) - R\sin(\alpha + \beta)]} \tag{2.7}$$

2.2.2.3 The Instantaneous Tangential Velocity of the Output Contact Point

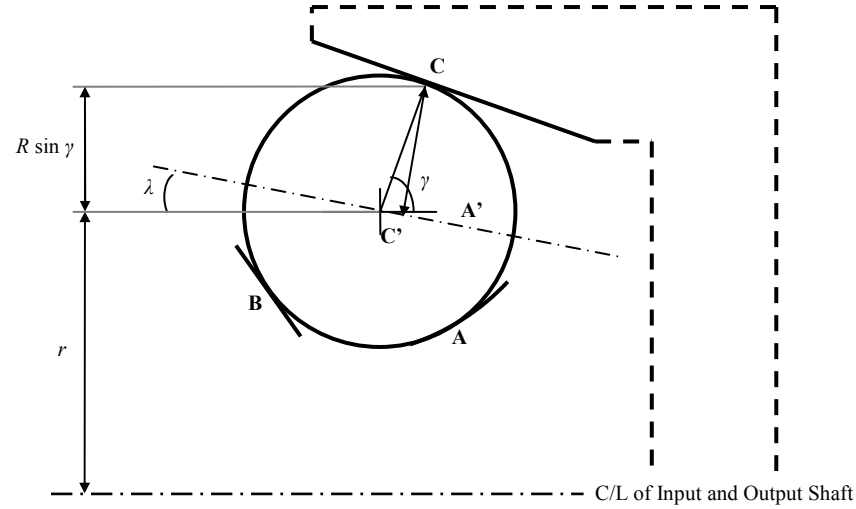


Figure 2-13: Calculation of Output Tangential Velocity

As before:

$$u_{ball-C} = \omega_{ball}[R\sin(\gamma + \lambda)] \tag{2.8}$$

Since the input and output shafts lay on the same axis, this also allows the calculation of the rotational speed of the output shaft (ω_{out}) from:

$$u_{out-C} = \omega_{out}(r + R\sin\gamma) \tag{2.9}$$

2.2.2.4 The Transmission Ratio

The transmission ratio '*i*' is simply a measure of the ratio of the input rotational speed to the output rotational speed, i.e.

$$i = \omega_{in} / \omega_{out} \tag{2.10}$$

Assuming $u_{in-A} = u_{ball-A}$ and combining Equations 2.2 and 2.4:

$$\omega_{in} = \frac{\omega_{ball}[R\sin(\alpha - \lambda)]}{r - R\sin\alpha} \tag{2.11}$$

Likewise, assuming $u_{ball-C} = u_{out-C}$ and combining Equations 2.8 and 2.9:

$$\omega_{out} = \frac{\omega_{ball}[R\sin(\gamma + \lambda)]}{r + R\sin\gamma} \tag{2.12}$$

Combining Equations 2.10 to 2.12:

$$i = \frac{\dfrac{\omega_{ball}[R\sin(\alpha-\lambda)]}{r - R\sin\alpha}}{\dfrac{\omega_{ball}[R\sin(\gamma+\lambda)]}{r + R\sin\gamma}} = \frac{\dfrac{\sin(\alpha-\lambda)}{r - R\sin\alpha}}{\dfrac{\sin(\gamma+\lambda)}{r + R\sin\gamma}} = \frac{\sin(\alpha-\lambda)(r + R\sin\gamma)}{\sin(\gamma+\lambda)(r - R\sin\alpha)} \tag{2.13}$$

In order to derive an equation for the transmission ratio as a function of only the five key dimensions and α, the Equations for r (2.1) and λ (2.7) must be combined with Equation 2.13. After simplification, this results in Equation 2.14.

$$i = \frac{\sin(\alpha+\beta)((r_0 - (R_1 - R)\sin\alpha) + R\sin\gamma)}{(r_0 - (R_1 - R)\sin\alpha)\cos\alpha\sin\gamma - R\sin(\alpha+\beta)\sin\gamma + (r_0 - (R_1 - R)\sin\alpha)(\sin(\gamma-\beta) + \sin\alpha\cos\gamma)} \tag{2.14}$$

2.2.3 Dynamics and Normal Forces

Looking at a single ball element:

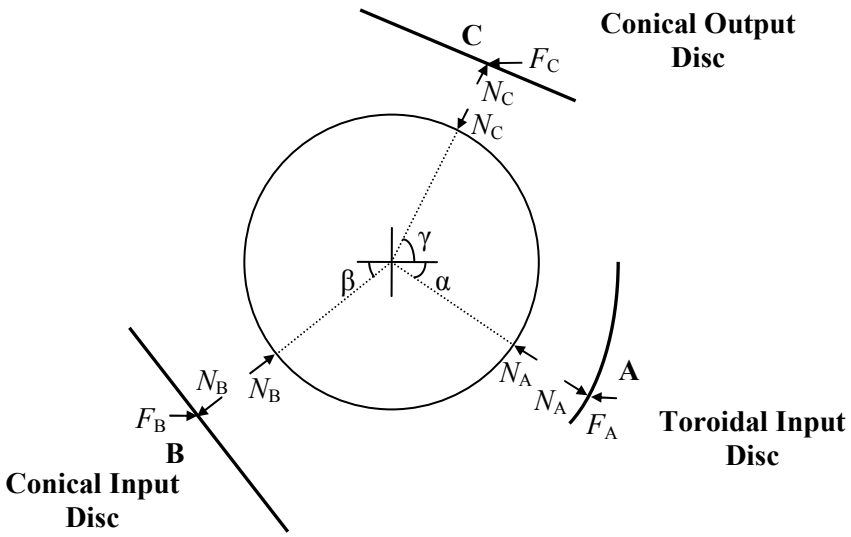


Figure 2-14: Free-body force analysis of single ball element

The normal forces applied to each contact point can be calculated from the free body force diagram shown in Figure 2-14.

Resolving:

$$N_B\cos\beta = N_A\cos\alpha + N_C\cos\gamma$$
$$N_B\sin\beta = -N_A\sin\alpha + N_C\sin\gamma$$

Combining:

$$\tan\beta = \frac{-N_A \sin\alpha + N_C \sin\gamma}{N_A \cos\alpha + N_C \cos\gamma}$$

Hence:

$$N_A = N_C \frac{(\sin\gamma - \cos\gamma \tan\beta)}{(\cos\alpha \tan\beta + \sin\alpha)} \tag{2.15}$$

Whilst a similar equation can be written for N_B:

$$N_B = N_C \frac{(\sin\gamma + \cos\gamma \tan\alpha)}{(\cos\beta \tan\alpha + \sin\beta)} \tag{2.16}$$

From Figure 2-14, the force produced by the ball screw coupling (F_C), can be related to N_C by the following equation:

$$F_C = N_C \cos\gamma \tag{2.17}$$

From the principle of a ball screw, (ignoring ball screw losses), the torque applied to the output (T_{out}) can be related to the axial force F_C as follows:

$$T_{out} = \frac{F_C l}{2\pi} \tag{2.18}$$

where l is the ball screw lead length, which must be chosen carefully to ensure that a sufficient axial force is applied to each contact to ensure that the traction coefficient (the ratio of tractive force to normal force) does not exceed the design value.

As before, the force applied to the toroidal disc (F_A) can be related to N_A by the following equation:

$$F_A = N_A \cos\alpha \tag{2.19}$$

The force F_A, which can either be produced by a simple spring or by a complex loading system, is crucial to the transmission ratio control of the CP-CVT. By combining Equations 2.15 and 2.17-2.19, the following equation is produced:

$$\frac{F_A}{\cos\alpha} = \frac{2\pi T_{out}}{l \cos\gamma} \frac{(\sin\gamma - \cos\gamma \tan\beta)}{(\cos\alpha \tan\beta + \sin\alpha)}$$

Rearranging:

$$T_{out} = F_A \frac{l}{2\pi} \frac{(\sin\alpha \cos\gamma + \cos\alpha \tan\beta \cos\gamma)}{(\cos\alpha \sin\gamma - \cos\alpha \tan\beta \cos\gamma)} \tag{2.20}$$

Hence the torque output is entirely dependent on the force applied to the toroidal disc.

By extension, combining Equations 2.14 and 2.20:

$$T_{in} = F_A \frac{l}{2\pi} \frac{(\sin\alpha\cos\gamma + \cos\alpha\tan\beta\cos\gamma)(r_0 - (R_1 - R)\sin\alpha)\cos\alpha\sin\gamma - R\sin(\alpha+\beta)\sin\gamma + (r_0 - (R_1 - R)\sin\alpha)(\sin(\gamma-\beta) + \sin\alpha\cos\gamma)}{(\cos\alpha\sin\gamma - \cos\alpha\tan\beta\cos\gamma)\sin(\alpha+\beta)((r_0 - (R_1 - R)\sin\alpha) + R\sin\gamma)} \quad (2.21)$$

2.2.3.1 The Concept of Constant-Power

The concept of constant-power may seem strange, since a transmission system's running power depends only on the power supplied to it (input torque and speed). However it can be stated that for an IC engine to remain in the same regime (and hence constant power output), the load applied to it must be constant. Equation 2.21 shows the dependence of the T_{in} (and hence engine load) on F_A, or in other words, there is a specific force that must be applied to the toroidal disc in order to maintain a constant torque input. Constant power can therefore be achieved by carefully controlling this force, either through a complex control mechanism, or ideally through the use of a simple spring.

Now it has been established that the CP-CVT is capable of supplying constant power, the question remains whether or not this is actually a desired feature for automotive applications. Given that a number of studies have looked in depth at CVT ratio control strategies it might seem idealistic to assume that the CP-CVT is capable of responding to driver demands automatically without additional control mechanisms. Whether or not this is feasible is discussed in more detail in Chapter 6, where the behaviour of the CP-CVT is simulated in a vehicular environment.

2.3 EHD Contact Behaviour

2.3.1 Fluid Rheology

An important aspect of any traction-type CVT is the EHD contact and traction fluid behaviour. The fluid acts as a lubricant and intermediary between contacting elements allowing the transferral of a shear force across the film without the surfaces contacting, thus reducing wear. Without the fluid, the surfaces would deteriorate extremely rapidly. This aspect of tribology has been extensively researched. Particular areas of interest for traction drives include fluid rheology/behaviour, contact losses and fatigue/failure.

One of the earliest papers to highlight the significance of fluid rheology in traction drives was the iconic work by Tevaarwerk and Johnson (1979). This paper is still considered relevant and applicable 30 years on, as indicated by the number of times it is referenced by a large proportion of modern papers regarding this subject. The paper comprehensively describes many of the primary aspects of traction fluid rheology from both a theoretical and practical approach, with specific reference to traction drives. The fundamental principle of traction drives are highlighted: “at sufficiently high shear stresses there is a marked nonlinear decrease in shear stress with rate” (Tevaarwerk and Johnson, 1979), whilst several graphs describe the effect of spin and side-slip on traction. Also described (albeit briefly) are spin-losses in traction drives, and a maximum efficiency operating point is offered (75% of its limiting traction), although this is explained more extensively in other papers.

This approach to traction was based around considering the EHD contact as a viscometer, where measurements of the geometry of the surfaces, the contact area and the friction force are used to determine values of average shear stress and average shear rate. This technique later enabled Evans & Johnson (1986) to identify the four distinct lubrication regimes according to the sliding conditions of the surfaces: linear viscous (Newtonian), non-linear viscous (Ree-Eyring), nonlinear visco-elastic and elastic-plastic. The theory and equations derived from the experiments conducted are based on some important assumptions, such as a shear flow across the contact, and that the shear stress at the wall of the contact is representative of the bulk property of the contact.

More recently Zhang, Zhang and Tobler (2000) developed a systematic model of a traction drive’s EHD contact based on the earlier work of Tevaarwerk and Johnson. This paper looks at a specific type of half-toroidal CVT with roller-type elements. An extensive set of equations is presented from first-principles, which concludes with a set of example calculations based on a specific component dimensions. Based on this model a graph is presented that shows the theoretical dependence of traction and side-slip forces on the roller axis offset. Despite the

inherent difference between the CP-CVT design, and the design discussed in this paper (rollers instead of ball elements, conical rather than toroidal output discs, etc), many of the methodology and calculations are universally applicable.

The calculation of film thickness as well as traction is often a complex process, as shown by both Chittenden et al. (1985) and Hirst and Moore (1978). Hirst and Moore (1978) used a two-disk machine to measure both film thickness and traction over a range of rolling and sliding speeds. This research was conducted before the development of modern traction fluids, and hence was conducted on simple mineral oils. They concluded that the film thickness given by earlier work (such as Dowson and Higginson, 1961) is still correct even at higher pressures, similar to the pressures encountered in modern traction drives. Sometime after this Chittenden et al. (1985) developed a number of equations and models based entirely on theory that agree well with previous research. Many of the equations developed are taken for granted within tribology today. One of these, which can calculate film thickness based on entrainment speeds, surface geometry and load, is used in later chapters (Equation 2.22).

$$\frac{h_c}{R_\sigma} = 3.0\left(\frac{2U\eta}{\overline{E}R_\sigma}\right)^{0.67}\left(\alpha\overline{E}\right)^{0.49}\left(\frac{2W}{\overline{E}R_\sigma^2}\right)^{-0.073} \tag{2.22}$$

More recently, Anghel, Glovnea and Spikes, (2004) provided experimental results for five traction fluids, looking in particular at the film behaviour and traction coefficient in various conditions. The tests were conducted on an industry standard mini-traction machine to measure traction, and an ultrathin-film interferometer test rig to measure film thickness (both manufactured by PCS Instruments). Rather than discuss the fundamental contact behaviour in detail, the paper instead presented a more straightforward and simpler approach to the calculation of film thickness based on simplified best-fit curves (Equation 2.23).

$$h_c = kU^a \tag{2.23}$$

where the values of k and a are given in the paper and vary depending on the specific traction fluid used, and lubricant temperature.

In addition to this simplified model of film thickness, the paper also presents detailed traction graphs as a function of slide-roll ratio (the ratio of the velocity difference in the contacting surfaces and the average surface velocity). From these graphs it is possible to extract best-fit curves for the traction coefficient, which for simple traction simulation is extremely useful.

The majority of these theories are based around, or incorporate the work of Hertz, and in particular the Hertzian theory of elastic contact, which was originally published by Hertz over a

century ago in 1882 (Johnson, 1985). Assuming elastic bodies in contact (which is often the case of steel components under typical loads) the deflection of the contacting surfaces can be determined from Hertzian theory. This theory for elliptical contacts (as can usually be found in traction drives) allows the calculation of the effective contact half-width (c), peak Hertzian pressure (p_0) and mean Hertzian pressure ($\bar{p}$), shown in Equations 2.24-2.26 respectively.

$$c = \left(\frac{3}{4} \frac{W R_\sigma}{\bar{E}} \right)^{1/3} F_1 \tag{2.24}$$

$$p_0 = \left(\frac{1}{\pi} \right) \left(\frac{6 W \bar{E}^2}{R_\sigma^{\ 2}} \right)^{1/3} F_1^{-2} \tag{2.25}$$

$$\bar{p} = \frac{W}{\pi a b} = \tfrac{2}{3} p_0 \tag{2.26}$$

In these equations, W is the normal load, R_σ is the reduced radii of curvature (which can be calculated from Equation 2.27), and $\bar{E}$ is the reduced contact modulus (Equation 2.28).

$$R_\sigma = \sqrt{R_x R_y} \tag{2.27}$$

Where:

$$\frac{1}{R_x} = \frac{1}{R_{x1}} + \frac{1}{R_{x2}} \quad \text{And} \quad \frac{1}{R_y} = \frac{1}{R_{y1}} + \frac{1}{R_{y2}}$$

Where convex surfaces curvatures are positive, whilst concave surfaces are negative.

And:

$$\bar{E} = \left(\frac{1 - \nu_1^2}{E_1} + \frac{1 - \nu_2^2}{E_2} \right)^{-1} \tag{2.28}$$

The other unknown in these equations is the variable F_1, which was originally presented graphically (Figure 2-15, from Johnson, 1985). In order to allow repetitive calculations of the Hertzian contact properties, a simplified, approximate relationship was developed that utilises a logarithm function, as shown in Equation 2.29. This equation demonstrated a very good correlation with the curve shown in Figure 2-15.

$$F_1 = -0.18 \log\left(R_x / R_y \right)^{1/2} \tag{2.29}$$

This theory of contact pressure and deformation generally dictate the fluid rheology and behaviour in traction drives, and hence is used significantly.

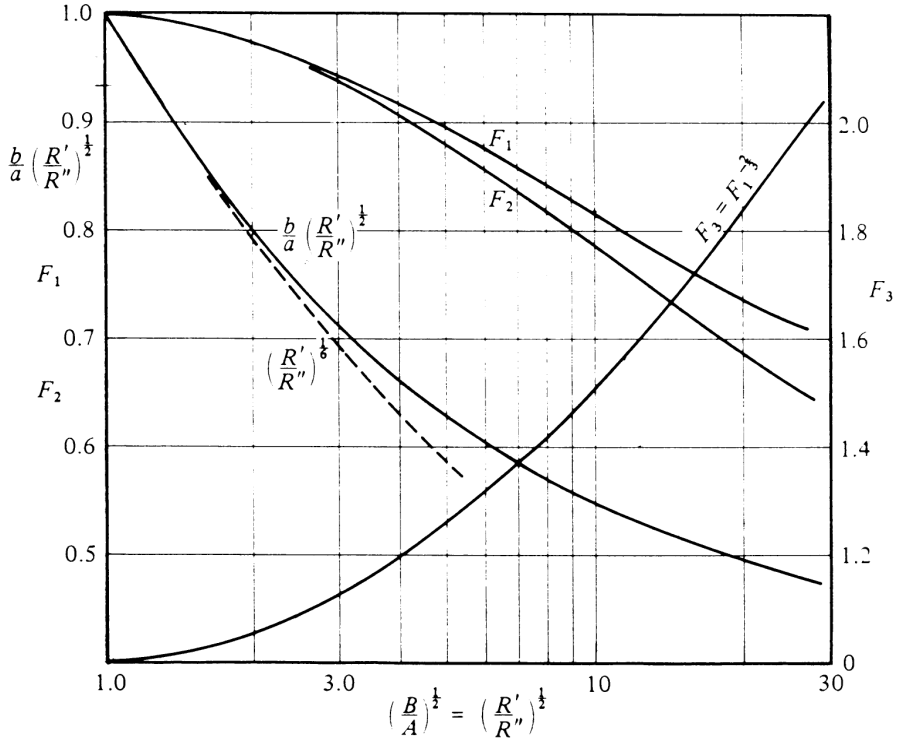


Figure 2-15: Relationship between F_1 and surface radii in Hertzian theory

2.3.2 Contact Fatigue

Contact fatigue can occasionally be overlooked in traction drives, however fatigue can and does still occur. One reason for this is the large tensile stresses that are present at the surface of the elements. Tanaka et al. (1995) explains that this effect can be reduced through the use of heat treatment, however this will not eliminate the problem entirely, partly because even with advancements in traction fluids, their viscosity is still largely affected by the high temperatures that occur in traction drives, reducing the film thickness. It must be highlighted however that this problem seems to have been somewhat solved in more recent years by further technological advances such as better surface finishes and hardening, and is no longer considered the primary cause of surface fatigue. Contact fatigue is strongly influenced by the presence of tractive forces (tangential stresses) on the contacting surfaces in addition to spin. This means there will always be at least some contact fatigue within traction drives, however by using ultra-clean bearing-quality steel, and taking measures to limit contact pressure, spin, and the traction coefficient, a traction drive's life can be significantly extended.

A comprehensive discussion of the fatigue life of traction devices was presented by Nikas (2002), who looked at the effect of several characteristics on the relative life of the contacting

elements. The relative life of each component is calculated theoretically, with specific reference to the following characteristics:

- Hertzian Pressure
- Slide-roll Ratio
- Roughness
- Traction Fluid Temperature

Once again, the differences in CVT designs means these results are not directly applicable, however the methodology used provides an extremely useful approach that could be used to predict the fatigue life of the CP-CVT. An alternative approach to fatigue is presented by Lee et al. (2004). Whilst Nikas (2002) used an entirely theoretical approach, this paper analyses fatigue using experimental validation. This approach includes the use of 3D surface texture measurements and specially designed software, providing possibly the most accurate assessment of fatigue. The conclusion however, is that "conventional stress and fatigue analysis methods remain applicable", hence validating the approaches shown in previous literature.

Generally, it is accepted that to ensure a reasonable operational life, the traction coefficient should not exceed 0.045 (Fuchs and Hasuda, 2004), and the Hertzian pressure must not exceed 2GPa (Lee et al., 2004). Given that traction fluid temperature affects the traction coefficient (Anghel, Glovnea and Spikes, 2004), and that surface roughness can be assumed to be similar to existing designs, these two restrictions should be sufficient to ensure a reasonable traction drive life.

2.3.3 Contact Losses

Within most traction drive EHD contacts there exists a number of distinct sources of loss (inefficiencies), including creep (slip), side-slip, and spin. Creep arises from the difference in the contacting surfaces' velocity that is required to produce the shear effect that produces traction, and hence acts in the primary direction of traction. Side-slip acts perpendicular to this and arises due to an offset in the relative axis' planes, and thus is not present in some traction drive designs; whilst in other designs (such as the Torotrak design) it is critical to the operation of the device as it forces the roller axis to rotate. Spin, which is perhaps the most difficult to visualise, acts around the contact's common normal, and hence the directions of traction, side slip and spin are mutually perpendicular. There are several ways of introducing the effect of spin and side-slip into the modelling of traction contacts: either the loss contribution of each aspect can be individually calculated, as shown later, or alternatively, more advanced traction

curves can be employed that incorporate these effects and determine their overall influence on the effective traction coefficient.

In addition to general literature that describes the losses that occur in EHD type contacts; there is also specific, more relevant literature that discusses these losses in relation to traction drives. For example, Yamamoto, Matsuda and Hibi (2001) analyse the overall efficiency of a half-toroidal CVT, and whilst the design of this traction drive is very different to the CP-CVT, several aspects (such as spin losses) are directly applicable. Unfortunately only a small section looks at methods for calculating losses. The remainder of the paper offers some suggestions for reducing losses in that particular design, and compares the calculated efficiencies with those determined using a prototype test-rig. A more detailed description and analysis of spin losses is presented by Sanda and Hayakawa, (2005). This paper refrains from looking at a specific traction drive design and instead offers a comparison of various layouts and how these layouts affect losses. A relatively simple test rig is also designed that provides experimental validation of the theoretically calculated losses. This paper also presents a very useful illustration of how spin arises and under what circumstances it exists (Figure 2-16, Sanda and Hayakawa, 2005).

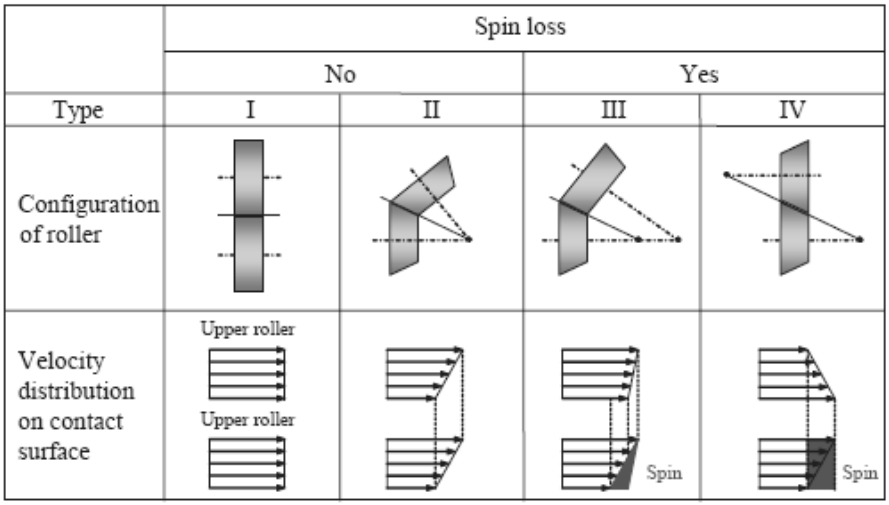


Figure 2-16: Occurrence of spin for different roller configurations

This paper uses a supposedly simplified model, which splits the contact into three distinct regimes in which only one of elastic, plastic or viscous-effects occur. Although this is a relatively simple approach compared to other models (which take a direct numerical solution of the full effects that occur within a contact) it still remains relatively complex, and generally beyond the scope of what is required for predicting spin losses. Despite this, a very useful graph is presented that directly shows the effect of spin on traction coefficient for various values of spin (Figure 2-17). This shows that as the spin velocity (ω_s) increases the effective traction

coefficient decreases. Or, put another way, for the same traction coefficient, increasing the spin velocity will increase slip and hence creep losses. These results are based on experimental rather than theoretical predictions.

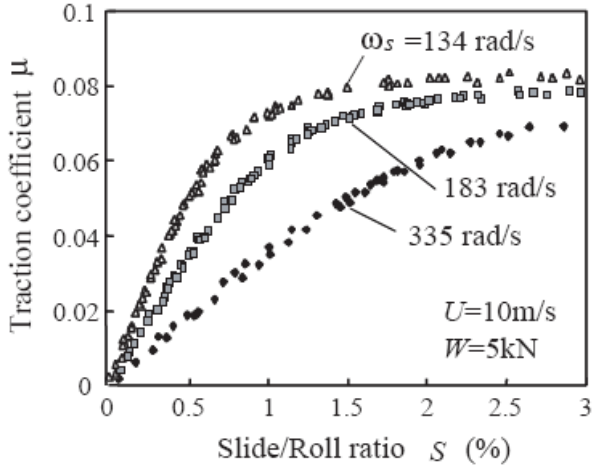


Figure 2-17: Effect of spin velocity on traction coefficient (Sanda and Hayakawa, 2005)

A different approach to spin losses is shown in Newall and Lee (2003), which again compares theoretical-derived models to experimental data. In addition to looking at the effect of spin on traction, this paper calculates the actual power lost to spin. Although these calculations make several assumptions, this approach is generally of more use for efficiency optimisation since the intricacies and mechanisms of spin are not particularly important, only a clearly-defined power loss. Unfortunately, this paper does not present, in any significant detail, the specific methodology used to predict spin, only stating that it is based on a loss factor, which itself is based on earlier work by Tevaarwerk and Johnson. Hence, whilst a good correlation is found between the model and experimentally obtained results, it is impossible to replicate the methodology used. The paper presented by Newall and Lee (from Torotrak), highlights how influential contact losses are on the overall efficiency of traction drives. Torotrak have invested significant resources into the analysis and reduction of contact losses in an attempt to improve contact efficiency, part of which is presented by Newall et al. (2004). This analysis begins with the assumption that was presented by Tevaarwerk and Johnson earlier: in order for the variator to be operating at its most efficient point, the force applied to the contacts should be controlled to give a traction coefficient of 75% of peak traction. However experimentation later found this assumption to be incorrect showing that optimum efficiency is actually achieved by operating at approximately 90% of peak traction. Whilst this figure is somewhat design-specific, and would not necessarily be corroborated for the CP-CVT, it does highlight the need for a full investigation into the specific contact efficiency for every unique variator design.

2.4 Other Traction Drive Losses

In addition to contact losses, a typical traction drive also has several other sources of inefficiency. In addition to high cage losses (which are design specific) the two main sources of loss in a traction drive typically come from the bearings that hold the input and output shafts, as well as the losses that are encountered as the rotating discs move through the traction fluid.

2.4.1 Bearing Losses

Bearings, in some form, are required for nearly all mechanical systems and hence a large amount of research and developments has been invested into reducing bearing losses to almost negligible levels. Despite this there is always a small amount of power lost to bearings in any mechanical system. These losses are typically defined in terms of a bearing operating torque, which is the amount of torque that opposes the direction of rotation. This torque depends on a several factors including rotational speed, the type of lubricant used, and the bearing's geometry. With the exception of heat lost to surroundings, the majority of the power lost in bearings directly goes to heating up the lubricant and the bearing itself.

Today most bearing manufacturers provide a certain amount of information regarding bearing losses, based on simple experimental setups. Historically, in a definitive paper, Witte (1973) attempted to derive a general theoretical expression for the operating torque of roller bearings, initially under thrust loads only, and then under both radial and thrust loads. One of the aims of this paper was to determine the heat generation caused by bearing friction in order to allow it to be compensated for. The paper begins by briefly discussing the test rig used to validate the equations produced, and then derives several equations based purely on correlation with the test results, with a relatively good correlation, as shown in Figure 2-18. The equations are hence not fundamental in nature and only serve to predict torque lost, rather than attempting to improve the efficiency of the bearings themselves.

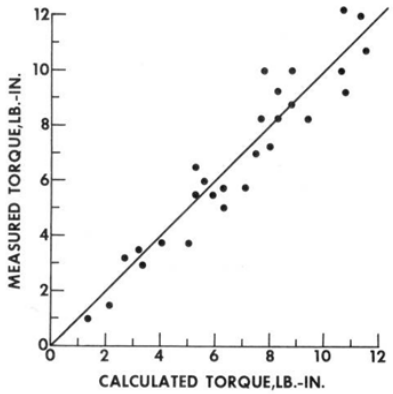


Figure 2-18: Correlation of measured and calculated torque (Witte, 1973)

A more recent and complete analysis of bearing torque is presented by Harris and Kotzalas (2007). This book summaries a wide variety of topics regarding bearing technology, including geometry, stresses, loads, and deformation. Of most use here is Chapter 10, which discusses kinematic speeds, frictional torque and power loss. A wide variety of bearing types are considered, of which the most applicable are the equations for tapered roller bearings. This type of bearing is capable of withstanding high axial loads coupled with radial loads, and hence is ideally suited for the CP-CVT. Although no information is given concerning how these equations have been derived or verified, they are clearly influenced by the earlier work of Witte (1979), and simply adapted for the change in units. Assuming a purely axial load, the bearing operational torque is given as:

$$M = 3.35 \times 10^{-8} G \sqrt{(n \nu_0)} F_a^{1/3} \tag{2.30}$$

where ν_0 is the lubricant kinematic viscosity; n is the rotational speed; F_a is the axial load and G is the geometry factor of the bearing, which can be calculated from Equation 2.31.

$$G = d_m^{1/2} D^{1/6} (Z \times l)^{2/3} (\sin \alpha)^{-1/3} \tag{2.31}$$

In this equation, d_m is the bearing pitch diameter, D is the roller diameter, Z is the number of rolling elements, l is the roller length, and α is the contact angle.

2.4.2 Churning Losses

Whilst bearing losses are relatively simple to calculate and there is general agreement within the research community, there still remains a large amount of contradiction in literature regarding the calculation and magnitude of churning losses. Churning losses are generally of more interest in dip-lubricated gears, and hence the majority of literature regarding churning looks at gear tooth geometry and how it affects losses. There is not a substantial amount of literature documenting the churning effect of smooth discs.

One of the earliest studies to look specifically at partially submerged smooth discs was conducted by Boness (1989). Boness used a specially designed test rig to measure the resistive torque due to churning of a variety of different discs and gears, varying factors such as diameter, speed, lubricant, depth of immersion and temperature. Based on this, he proposed a series of equations that show the churning torque as a function of a moment coefficient (C_M), which itself is a function of the Reynolds number of the interface between the rotating surface and the lubricant bath. Later on Terekhov (1991), used a similar methodology to find the churning loss (or specifically, heat generation) of gears rotating in oil, and derived similar equations based on a moment coefficient and Reynolds number. A more recent study (Luke and Olver, 1999)

compared the results produced by both of these models to experimental results derived from an updated test rig (albeit with gears only). A comparison of the two models is shown in Figure 2-19 (Luke and Olver, 1999).

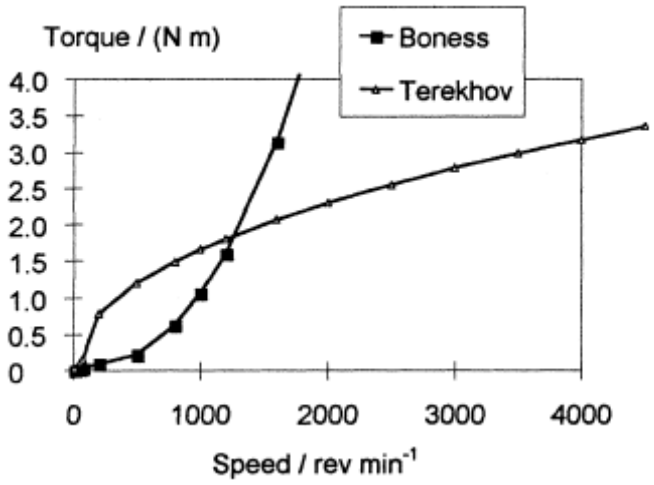


Figure 2-19: Comparison of Boness' and Terekhov's churning torque models

This figure revealed that Boness' original model vastly overestimated churning torque at higher rotating speeds. However, Luke and Olver also claimed that Boness' model is in fact more accurate, provided a modification is made to the calculation of C_M, and the Reynolds number. A comparison of the differences (assuming turbulent flow) is shown in Table 4.

Table 4: Luke and Olver's modified equations for churning torque calculation

	Boness' Original Equations	*Luke and Olver's Modified Equations*
Moment Coefficient	$C_M = (5\times10^8)\,\mathrm{Re}^{-2}$	$C_M = (5.34\times10^4)\,\mathrm{Re}^{-1.379}$
Reynolds number	$\mathrm{Re} = \frac{\omega RL}{\nu_{oil}}$	$\mathrm{Re} = \frac{\omega RL}{\nu_{air}}$

Given that CVT input shafts typically rotate at the engine speed (1000-6000+ rpm), the use of Boness' original equations would yield unrealistically high torque values, hence the findings of Luke and Olver must be incorporated.

2.5 Design Techniques

2.5.1 Theory of Design

2.5.1.1 Customer Requirements

One possible method of improving the CP-CVT further is to utilise and adapt a number of existing design theories. Several design theories exist intended to aid development during the design stage of a wide variety of engineering problems. One such theory is Quality Function Deployment (QFD), which attempts to "satisfy customers by translating their demands into design targets and quality assurance points" (Akao, 1990). This initially requires the knowledge of customer demands, which, without specific market research, must be taken from existing literature. One of the earliest literatures to look at transmission development from a customer-orientated approach was Baudoin (1979), whom stated that there are 6 key requirements that must be considered by any transmission designer: Mass producibility at low costs; High efficiency (ability to give a good fuel economy); Comfort (low noise and limited 'jerk' when changing gears); Limited weight and volume; Limited maintenance; Good reliability and life.

Sometime later, in a book specific to traction drive design selection, Heilich and Shube (1983) provided detailed information regarding the selection of traction drives including an outline of a detailed specification sheet that should be included with all traction drive designs. This provides an excellent overview of the typical factors that should be considered in transmission design. These factors are divided into eight categories, which are summarised below:

1. Input: Power rating, speed, torque, input shaft size, rotational direction, method of input shaft coupling.
2. Output: Speed range required, torque at maximum and minimum speed, output shaft size, method of output shaft coupling.
3. Adjustment: Local/remote, method of adjustment (mechanical, electrical, hydraulic, etc)
4. Mounting: Horizontal or vertical, orientation of input and output shaft.
5. Load: Inertia, shock loading, restarts and reversals per hour, speed adjustments required, speed-holding precision, time spent in each ratio.
6. Operating conditions: temperature, humidity, hazards (dust, chemicals, etc).
7. Access: regular access, removal of motor, inspection and replacement of lubricant
8. Power source: type and control method.

In addition to providing a history of various traction drive designs, Heilich and Shube (1983) also summarise in terms of cost, performance and application, all the current designs. Unfortunately this is largely outdated now due to significant technological advances since publication and such an in-depth assessment of traction drive design has not been replicated more recently.

More recently, in the book "Rotary Power Transmission Design", Hurst (1994) provided a good overview of the process of transmission design, stating the typical factors that influence gearbox design and choice. These factors are divided into five categories: commercial, performance, installation, operation, and environment. A summary of the factors stated by Hurst (1994) is shown in Table 5. Since these requirements are aimed at more general applications rather than specifically for automotive purposes, not all of them are relevant. The requirements have thus been divided into three categories, indicated by the colour shown in the table:

1. Not-applicable (green): Design considerations, such as appearance, or altitude, which clearly do not apply to automotive transmissions
2. Assumed Acceptable (blue): If it is assumed that the transmission design will not vary significantly from existing designs, then a number of considerations do not have to be actively considered
3. Critical (red): Design considerations that must be improved, or at least considered at all stages of design

Table 5: Summary of typical factors affecting gearbox selection

Commercial	*Performance*	*Installation*	*Operation*	*Environment*
Quantity required	Power required	Relationship in system	Reliability	Space available
Price	Maximum Torque and speed	Shaft orientation	Maintenance	Size
Commercial life	Speed ratio	Ease of installation	Accessibility	Shape
Cost	Duty cycle	Weight	Consequences of shut-down	Load capacity of supports
Convenience in purchase	Mechanical integrity	Mounting type	Condition monitoring	Ambient conditions
Manufacture reputation	Radial/Axial and shock loads	Positional location accuracy	Maximum operating temperature	Altitude
Appearance	Inertia, efficiency	Support loads		Heat dissipation required

Based on the works of Baudoin (1979), Hurst (1994) and Heilich and Shube (1983), and literature discussed previously, a number of key customer requirements have been extracted: low noise, reduced fuel consumption, ease of implementation, performance (power capacity), drivability (fast response), reliability/durability, and low capital cost.

2.5.1.2 Design Processes and Tools

According to Hurst (1994), which itself is based on the earlier work of Pugh (1986), the design selection and development procedure can be divided into five main areas: market, specification, concept design, detail design, and manufacture, as shown in Figure 2-20.

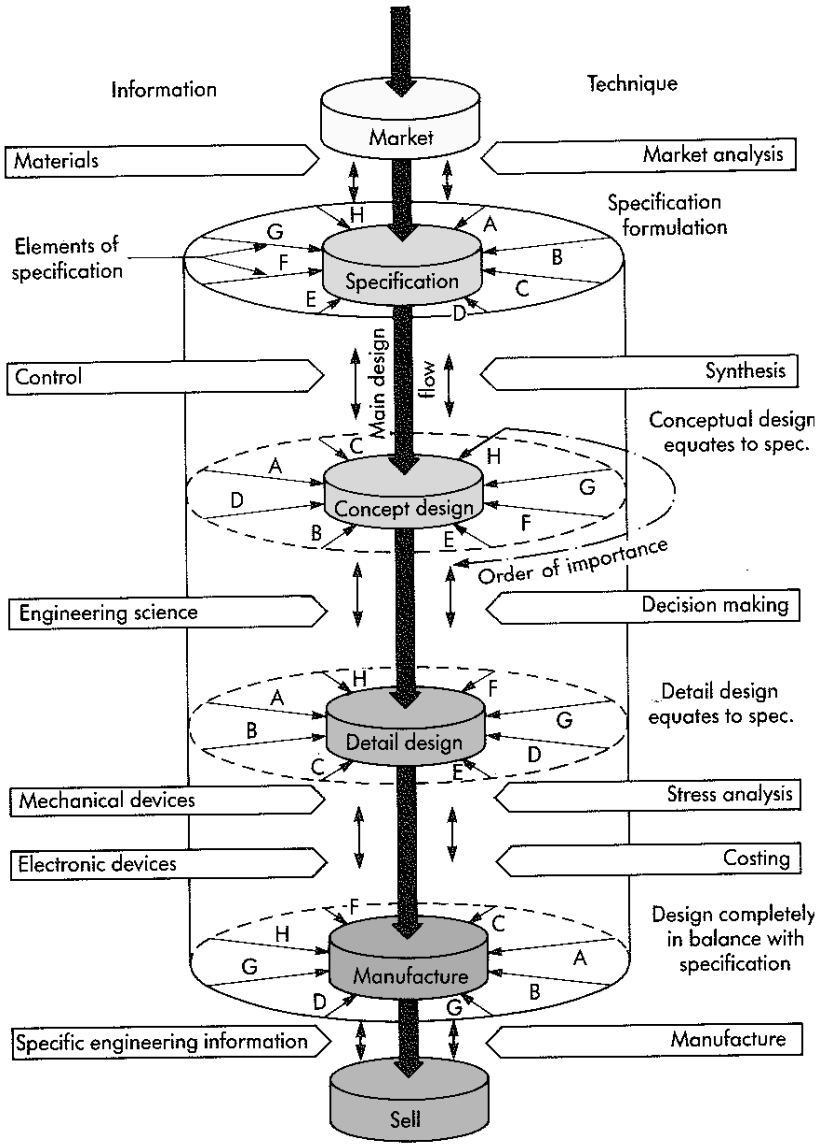


Figure 2-20: Design activity model (Hurst, 1994)

The literature discussed in the previous chapter illustrates there is definitely a **market** need for the CP-CVT in order to improve on existing designs and offer the customer improved efficiency and simplified control. The **specification** (in terms of customer requirements) has been derived from the works of Baudoin (1979), Hurst (1994) and Heilich and Shube (1983), albeit without specific targets. Furthermore, the initial **concept design** of the CP-CVT design has already been established. Hence, the focus must now be on the **detail design** and final concept decision

making, in order to bring the CP-CVT to a stage where it is ready to be manufactured and implemented. In order to achieve this, there are a number of tools available.

One of these tools is Quality Function Deployment (QFD), mentioned previously. This process, which was originally proposed by Akao in Japan in 1966 (Akao, 1990) has developed into a mechanism to translate the voice of customers into engineering language. According to Gover (1996), QFD is more of a process than just a tool for product development based on the concept of "Company Wide Quality Control". Its essential characteristics are considered to be customer-orientation and team-approach aimed at concisely structuring communications and linking information together. It can be adapted to include the use of Pugh's concept selection method (Pugh, 1981) and a translation chart, known as the 'House of Quality' (Houser and Clausing, 1988).

In the iconic work "Concept selection: a method that works", Pugh (1981) developed a now widely used method to aid in decision making in design. It is based around the use of an evaluation matrix, which is associated with the QFD method, and is essentially a method of prioritising different criteria. The process involves constructing a matrix that rates certain evaluation criteria against various concepts. A relative 'score' for each concept can be created through the use of a baseline score from which each of the concepts is rated against. Originally this scoring was done using symbols (which later influenced the use of symbols in the House of Quality) with each symbol indicating either a positive, negative or neutral scoring, although more recently numbers are used instead. Criteria scores are then combined to give a total score for each concept, with larger scores indicating a concept is preferable in terms of the criteria used. According to Pugh (1981), this method is effective for comparing alternative concepts and is an iterative method that can be incorporated at each design-decision stage.

Pugh later focused more on the complete design process and model, including creating the design model shown in Figure 2-20. Pugh's ethos was based around several criteria regarding any design processes (Pugh, 1986): all must be able to relate to it; all must be able to understand it; all must be able to practice more effectively and efficiently as a result of using it; it must be comprehensive; it should have universal application.

Perhaps the most successful methodology to arise out of this philosophy is the House of Quality (HoQ), which is designed to be easily understood by all, whilst still containing all necessary and relevant information. Perhaps the best way of describing the House of Quality was offered by Houser and Clausing (1988), whom asked the question: "design is a team effort, but how do marketing and engineering talk to each other?" Hence the HoQ is simply a tool that allows the

intricate relationship between customers' demands and engineering requirements to be determined., and includes information on 'what to do' (customer requirements), 'how to do it' (engineering requirements) (Bai and Kwong, 2003). An example of this is shown in Figure 2-21. Although at first glance this might seem quite complicated, it can be simplified by dividing it into a number of separate parts, as shown in Figure 2-22.

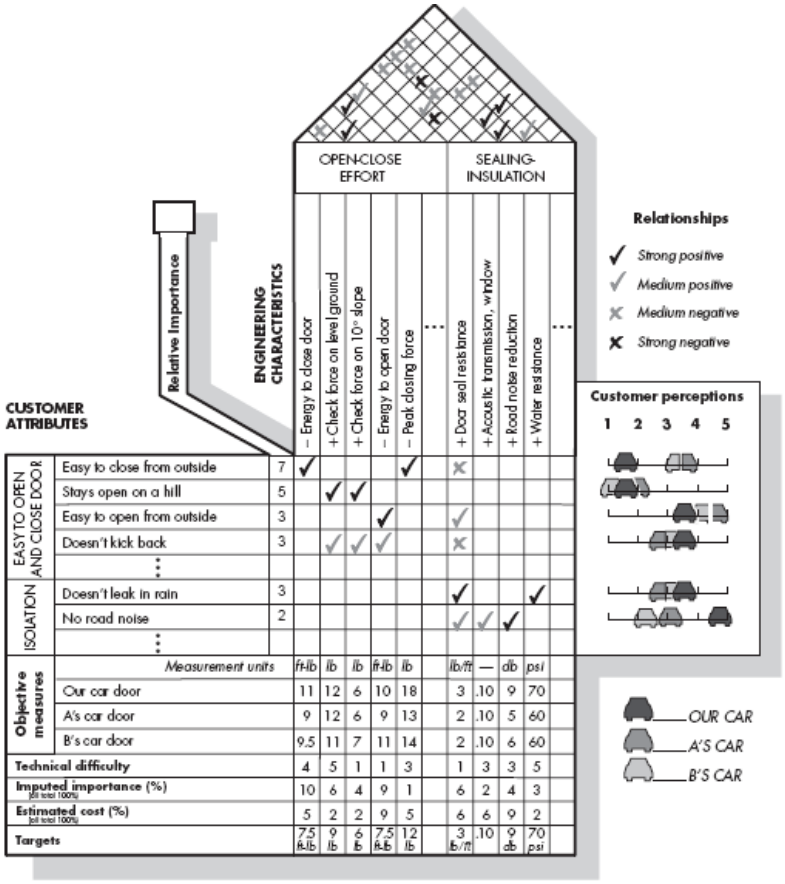


Figure 2-21: Complete House of Quality (Houser and Clausing, 1988)

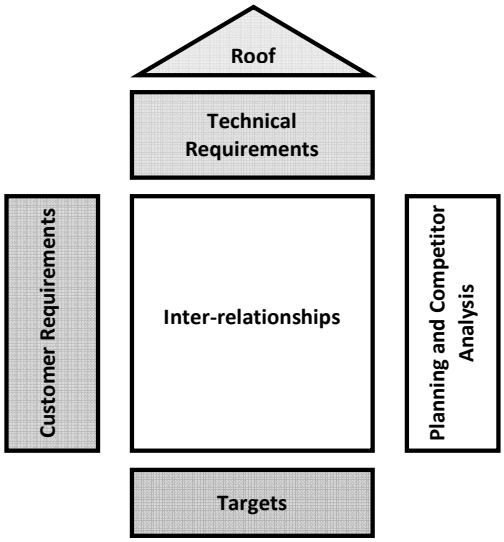


Figure 2-22: Aspects of the House of Quality

The **customer requirements** shown in the HoQ are the demands determined through market analysis that describe the effectiveness of any design in fulfilling customer satisfaction. These are related to specific, measurable **technical requirements**, which can be determined for any particular concept or design choice. The **inter-relationship** matrix, based on the earlier work of Pugh (1981), shows how much each of the technical requirements influences customer satisfaction. This is typically done either symbolically, with symbols representing either a strong, medium or weak effect, or numerically. The debate about which type is superior (symbols, or numbers) was discussed by Merts (2009), whom stated that neither format is universally superior to the other. On the contrary, they both serve different purposes and are uniquely suited for working with different groups: symbols provide more visual impact, useful for marketing audiences, whilst numbers provide a more technical appearance, better suited for engineers and scientists. The **planning and competitor analysis** section looks at how customers perceive each of the requirements in the current design compared to competitors' designs, again this can either be numeric or symbolic. The **roof** is used to demonstrate how changing one technical requirement will affect the others in either a positive or negative way. Finally the **targets** are shown at the bottom, which are usually based around comparison to competitors. Advanced HoQs also employ derived functions that rate the importance, difficulty and cost of each technical requirement to provide more information to the engineer when determining whether design solutions are appropriate or not. The HoQ will not yield specific targets; only demonstrate how important they are in satisfying customers.

It the determination of the target values for engineering requirements that is perhaps the most difficult process, hence methods have arisen that attempt to simplify this process. These methods are normally based around the use of fuzzy values, either with a genetic algorithm solver (Bai and Kwong, 2003) or fuzzy-regression (Kim et al., 2000). The aim of these processes is to determine targets that yield the highest customer satisfaction.

Kim et al. (2000) used a rather complex process to develop an integrated formulation and solution approach aimed at helping the designed to fill in certain aspect of the HoQ (namely, system parameters, objectives, and constraints). The methods used included complicated multi-attribute value theory combined with fuzzy regression and fuzzy optimisation theory. The purpose was to seek the highest overall customer satisfaction by removing elements of subjectivity from the HoQ process. It is also claimed that this methodology can serve as a guideline to determine how much flexibility is warranted or possible in design decisions (Kim et al., 2000).

Bai and Kwong, (2003) focused more on determining target values of the engineering requirements of the HoQ. Using the car-door example from the original HoQ paper (Houser and Clausing, 1988), they developed an in-exact optimisation model intended to provide preferred target values. Rather than obtaining a single set of exact targets, the methodology generates a combination of preferred solution sets (all with acceptable customer satisfaction) from which an exact set of target values can be produced based on the specific design scenario.

2.5.2 Optimisation Algorithms

Assuming target values can be obtained, the next process in design is determining how these target values can be fulfilled. A number of optimisation techniques exist, including evolutionary, brute force, particle swarm, hill-climbing and iterative algorithms. Both brute force (which involves the calculation of all possible solutions) and iterative (which involves the use of steadily increasing iteration to evaluate specific solutions of interest) are very simple and hence are not generally discussed in literature. Likewise, hill-climbing, which simply involves adjusting each input variable systematically to determine better adjacent solutions, is straightforward enough that it doesn't warrant discussion in literature. More advanced techniques such evolutionary algorithms and particle swarm optimisation, are far more useful in practice, designed to efficiently determine optimum solutions based on exact, or in some cases fuzzy, targets.

2.5.2.1 Genetic Algorithms

There is a plethora of research conducted into evolutionary algorithms, and especially into genetic algorithms. These algorithms are designed to mimic the process of natural evolution, and in particular, natural selection. A good overview of this research is presented by Whitley, (2001), which summarises the processes involved in genetic algorithms, and discusses the controllable attributes involved, such as mutation and cross-over, and how they affect the behaviour of the algorithm. According to this paper, genetic algorithms were developed in the United States in the 60s and 70s and are designed for use in several different disciplines, such as search, optimisation, machine learning, and of most interest here, solving design problems. Historically, it is claimed, genetic algorithms and evolutionary strategies are distinctly different, with the former placing more emphasis on selection criteria, and recombination over mutation, and the latter placing more emphasis on mutation over recombination. Additionally, genetic algorithms tend to encode variables into strings of binary bits, which together form a 'chromosome', whilst evolutionary algorithms tend to use more direct representation of the variables. More recently however, these terms are largely interchangeable as the fundamental

traditions have somewhat merged. Both techniques follow a fundamentally similar routine (adapted from Whitley, 2001; Holland, 1992; and Biles, 1994):

1. Initially a number of random solutions are generated and evaluated, forming an initial gene-pool (0^{th} generation).
2. Based on the evaluated solutions, each member (chromosome) of the gene pool is assigned a 'fitness' score based on how close the solution is to the desired result.
3. An 'intermediate generation' is then created, typically using roulette (or similar) selection based on each chromosome's score. Fitter chromosomes (one that scores higher) have more chance of being taken into this intermediate generation.
4. Pairs of chromosomes are then randomly selected from this intermediate generation and then crossed-over to create the next generation. Cross-over involves taking part of one chromosome and combining it with part of another.
5. This process is then repeated from step 2 as necessary.

Additional factors that influence the way the algorithm operates are cross-over rate, rate of mutation, population size, and number of generations. The selection of appropriate population size and the number of evolutions is generally problem-dependent. This together with the string length (which is determined by the resolution of solution required), have a large influence on the overall computational time required. Certain problems for which accuracy is less important may choose to use shorter bit-string lengths, and fewer evolutions, whilst those who require consistent, accurate results, may choose to use large population sizes and string lengths and run more evolutions. The problem with this approach is that the algorithm may start to converge to false local maxima, which is one of the reasons random mutations are used. Random mutations essentially flip a single bit within a chromosome, changing the value of a particular input variable by a small amount, ensuring the no particular solution or area is ever fully abandoned. Determining the mutation rate (probability of mutation) is relatively important; if it is set too high the natural evolutionary process will be detrimentally affected, whilst if it is too low, certain possible solutions will remain eternally unresolved. Typically the mutation rate for genetic algorithms is less than 1-2% (Whitley, 2001).

Cross-over is the process of combining two chromosomes from a generation to populate the proceeding generation. The cross-over rate determines the probability of chromosomes being combined. Higher cross-over rates mean that certain good solutions maybe lost by being combined with poorer solutions, whilst lower cross-over rates mean that solutions are carried forward un-evolved, meaning more evolutions are required to produce better solutions. Another important factor is the cross-over point, which can either be single, two, or multi-point. In single-point combinations, two chromosomes are combined at a random point, to produce two

new chromosomes. In two-point combination the chromosomes are crossed over twice, and in multi-point they are crossed over many times. This is illustrated in Figure 2-23.

Original Chromosome (Parents)	Chromosome 1	xyxyxyxyxyxyx
	Chromosome 2	abababababababa
Single-Point Cross-over	Chromosome 1	xyxyxyx bababa
	Chromosome 2	abababa yxyxyx
Two-Point Cross-over	Chromosome 1	xyxy ababab xyx
	Chromosome 2	abab xyxyxy aba
Multi-Point Cross-over	Chromosome 1	ab x ba xy a y ab xy
	Chromosome 2	xy a yx ab x b xy ab

Figure 2-23: Examples of single, two and multi-point cross-over

Which cross-over method is best has been the subject of a lot of research in itself (Syswerda, 1989; Caruana et al. 1989; De Jong, 1975). Since chromosome strings are simply concatenated strings of bits representing each input variable, the use of single or two-point crossover will change at most one or two of the variables, with the others remaining the same but being combined with another set of variables. This means that there is less mutation and evolution of the variables, but conversely less chance of destruction of potentially strong chromosomes that can arise from multi-point crossover (Fonseca and Fleming, 1995).

The crossover and mutation rates largely determine the accuracy of the solution offered and furthermore the speed at which the solution convergences. Recently, more advanced adaptive genetic algorithms (AGAs), use advanced techniques to determine suitable crossover and mutation rates and automatically based on the current population information (Srinivas and Patnaik, 1994). These techniques are extremely difficult to implement and require a very good understanding of the nature of the problem, the search space, and the relative fitness of the current population. These techniques go beyond the scope of this work.

The advantage of using genetic algorithms, as stated by one of the original developers (Holland 1992), is that each evolution is in effect evaluating a number of parallel solutions simultaneously. In order to better demonstrate this point, Holland proposes that a problem with n number of input variables can be viewed as an n-dimensional hypercube, with each point in the hypercube representing a single possible solution. Each time a solution is evaluated, it is

implicitly evaluating solutions that run along the same hyper-plane. He argues therefore that far more hyper-planes are sampled than the number of evaluations that occur in each generation.

Whether or not this is strictly true depends somewhat on the nature of the problem. This view of the problem as a search space implies that genetic algorithms should not necessarily be thought of as *optimisation* techniques but more *search* techniques, a view supported by both Holland (1992), and De Jong (1975). Each new generation can therefore be seen to simply sample many solutions from the search space, evaluate them, and then direct the search accordingly (Wright et al. 2005).

Genetic algorithms are not universally accepted or praised, with certain sceptics arguing that they are far too computationally demanding, being forced to evaluate nonsensical solutions, and that for certain, specific problems other optimisation algorithms may find better solutions than genetic algorithms given the same computational time. In any case evolutionary algorithms remain a very popular method of optimisation from applications involving game theory, design optimisation, artificial intelligence and even jazz improvisation (Biles, 1994).

One clear advantage of genetic algorithms is the ability to fulfil multiple criteria, as shown by Fonseca and Fleming (1995), who use Pareto optimality to optimise many objectives simultaneously. The purpose of Pareto optimality is to determine a situation in which it is impossible to make one criteria better without necessarily making another worse. For example in the context of optimising a concept design it maybe possible to improve the efficiency of a device, but at the cost of increasing the mass. Based on the relative importances of each criterion, and their target values, it is possible through the use of multi-criteria genetic optimisation to determine if this is valid.

Additional research is focused around the removal of the need for fitness evaluation. One particular method of achieving this is the use of “tournament selection” to select fitter chromosomes (Andrzej and Stanislaw, 2000). In this case rather than comparing solutions to a global target, they are instead compared to one another, with the stronger of the two progressing to the subsequent generation. This paper arranges the algorithm tournament in such a way that only feasible solutions are evaluated, which is a big advantage in certain optimisation problems, such as dimensional optimisation, since these tend to produce a significant number of nonsensical solutions. By eliminating the need to evaluate these, a considerable amount of computational time can be saved, which is confirmed in the results of this paper when compared to a more traditional approach. It is therefore surprising that this method of fitness evaluation has not been more widely used. One possible reason for this is that by removing global fitness

evaluation, it is only possible to determine the relative strength of each solution rather than the absolute strength compared to a predetermined target.

2.5.2.2 Swarm Optimisation Algorithms

An alternative optimisation technique to genetic evolution is the use of swarm optimisation of which several techniques exist, including Gravitational Search Algorithm (GSA; Rashedi, Nezamabadi-pour and Saryazdi, 2009.), Particle Swarm Optimisation (PSO; Kennedy and Eberhart, 1995), Ant Colony Optimisation (ACO; Dorigo, 1992), and Intelligent Water Drops (IWD; Shah-Hosseini, 2008). Many of these techniques are so similar in nature to one another that it is questionable whether they are indeed unique algorithms, and not purely the result of poor literature reviews by the authors involved. Perhaps the most widely discussed and implemented swarm techniques are particle swarm optimisation and ant colony optimisation, both of which have been applied to a variety of multi-criteria optimisation problems.

Like evolutionary algorithms, particle swarm optimisation is a search technique designed to find the optimum solution within a search space. The basic premise of particle swarm optimisation, originally proposed by Kennedy and Eberhart (1995) is a search method that optimises a problem by continuously improving a candidate solution with respect to a given scoring function. Optimisation occurs by having a population of candidate solutions (particles), which move through the search-space according to simple function. This function takes into account all the information that has been learned about the search space thus far, and hence particles will tend to gather (swarm) around particular regions of interest within the search space, indicating the best solutions are found there. The advantage of this technique is that requires far less computational effort to implement compared to evolutionary algorithms, and also requires no assumptions to be made about the nature of the search space itself. Specifically, particle swarm optimisation does not require the optimisation problem to be differentiable, which can be required for more complex or multi-criteria evaluation functions where a representative function is preferable. Furthermore, since this technique can be applied continuously using the evaluation functions directly it can be applied to irregular optimisation problems or ones that change with respect to time (Kennedy and Eberhart, 1995). Despite its simplicity (or perhaps because of it), particle swarm optimisation does not attract as much research literature as more advanced optimisation techniques. One of the biggest perceived flaws with this technique is the problem of premature halting, which can occur when particles swarm to local false-maxima.

A similar optimisation technique, Ant colony optimisation (ACO), was originally proposed in 1992 to determine an optimal path across a graph, based loosely on the behaviour of ants

(Dorigo, 1992). In order to explain this concept further, it is necessary to discuss briefly the evolved behaviour of ants. When beginning to set up home at a new location, ants begin by roaming the surrounding area at random in search of food. When they find food they return to their home laying down a trail of pheromones. When further ants find a trail like this they tend to follow it to the source of food, and when returning they strengthen the pheromone trail further. Over time, longer pheromone trails tend to fade, whilst shorter ones, which are continuously reinforced become stronger, and eventually all the ants will follow a single, efficient trail (Goss et al., 1989). The ACO algorithm is designed to essentially imitate this behaviour. Although it was originally intended to be used for efficient path finding, it has since evolved and is now applied to wider variety of problems including routing, assignment, scheduling, and multi-objective optimisation (Dorigo, Birattari, Stutzle, 2006).

A new swarm optimisation technique is proposed in a later chapter that uses aspects from both PSO and ACO to efficiently and quickly determine a fuzzy region with a search space that will contain the best solutions to a particular optimisation problem.

2.6 Literature Review Discussion and Summary

2.6.1 Transmission Technology

A number of different transmission technologies have been discussed, including standard step-gear transmissions, both automatic and manual, and more recent CVT designs. Manual transmissions remain the most popular in Europe due to their better performance, higher fuel economy, and increased perceived control. Outside of Europe, automatic transmissions tend to be favoured due to their ease of use, despite their inherent disadvantages. Both of these existing technologies could potentially be replaced by dual clutch technology (DCT), which is essentially the best of both worlds, incorporating the performance and efficiency of manual transmissions with the ease of use of automatics. This is achieved through the use of a hydraulically actuated clutch rather than a torque convertor, and can be run in either full or semi-automatic modes. All of these existing technologies still suffer from the same problem however, which is the use of discrete gear ratios. An engine is a complex machine that has a near-infinite number of operating points. The use of the correct operating point to produce the demanded torque and power as efficiently as possible is essential to improving the overall fuel economy of an automotive vehicle. This can only be achieved some of the time by discrete ratio gear boxes. In order to improve fuel efficiency and performance further, a CVT is required. A CVT, which has a continuous range of ratios available, allows a change in torque and speed at the wheels, whilst maintaining a higher engine efficiency operating point more often.

A number of different CVT designs exist and have existed for over a century. CVTs have traditionally been dismissed due to their limited torque and durability, however these problems have been solved in recent years, and the race is currently on to produce the CVT design that will be widely adopted, in particular by the automotive industry. The design with the most promise is the toroidal traction-type CVT, which utilises a special traction fluid in order to transmit shear force from one part to the next. A number of different traction designs have been developed, but have not been particularly successful due to the complex control mechanism required for synchronisation and ratio selection. Any design that hopes to be successful must hence address these issues as a priority. One such design, the CP-CVT, appears to meet these criteria, and furthermore, has been approved for patenting through an external company report (Nerac) which specialises in assessing the intellectual property potential in novel designs.

2.6.2 The Constant-Power CVT

Given the rather specific and specialised nature of the CP-CVT design, it is not surprising that there is only a small amount of literature available. The papers discussed generally serve as a

method of introducing the design to the research community, with each paper adding an additional element such as superficial geometrical optimisation or the inclusion of an additional set of ball elements (cage). The remainder of this subsection shows a kinematic and dynamic analysis of the design, based on fundamental theories and inspired by the literature discussed. Essential equations are derived for the transmission ratio, torque output, and normal forces, which are used continuously throughout this thesis.

The equations shown demonstrate that the CP-CVT can be controlled automatically based on the resistive torque (load) applied to the output shaft, and furthermore that a constant power output is theoretically possible. Whether this is a desirable feature is discussed in more detail in later chapters.

2.6.3 EHD Contact Behaviour

It has been established that the general behaviour of elastohydrodynamic lubricants is relatively well understood from both a theoretical and experimental approach. Specific aspects such as fluid rheology, contact fatigue and contact losses have been discussed, with relevant models and equations from existing literature also presented. The majority of this work is based around the initial work of Hertz, and more recently Tevaarwerk and Johnson, whose equations form the basis of nearly all traction and general tribological calculations.

It has been established that contact fatigue is influenced by a number of characteristics including Hertzian contact pressure, slide-roll ratio, surface roughness and lubricant temperature. In order to ensure a reasonable traction drive life two primary restrictions must be observed: the traction coefficient of the fluid should not exceed 0.045, whilst the Hertzian pressure should not exceed 2GPa.

In the majority of traction drives there are three main sources of contact loss: spin, side-slip and creep (tangential slip). Side-slip is only present when there is an offset between the planes of the axis of rotations of the contacting elements, and hence is not present in the CP-CVT. The calculation of creep losses is fairly trivial based on the extraction of the slide-roll ratio, whilst the calculations for the spin losses are far more convoluted requiring the contact parameters to be known. Hence the majority of the properties of a traction drive including wear, traction, fatigue and losses are dominated by the behaviour and nature of the contact, such as its thickness, dimensions and pressure, all of which are still determined from the work of Hertz over a century ago.

2.6.4 Traction Drive Losses

In addition to tribological/contact losses, there are also a number of other sources of loss within a traction drive. These include losses due to the cage (ball separator), bearings and lubricant churning. Ball separator losses are fairly specific to the CP-CVT design and are discussed later. Both bearing and churning losses can be expressed as a value of torque lost. The value of this torque can be calculated based on a number of literature sources, as discussed.

Despite their relatively small effect, bearing losses are generally better understood, perhaps because bearings are used in nearly all mechanical systems. Most bearing manufactures provide some information regarding the magnitude of the bearing operational torque, although for the majority of applications this is rather irrelevant since the losses are so small that they hardly affect the operation of a device. When attempting to accurately determine the overall efficiency of a device such as a traction drive however, these losses are important and must be considered, hence a number of equations are presented to calculate the torque lost to bearings based on existing literature.

Conversely churning losses can be relatively high, and yet there is not a large amount of literature regarding the magnitude and calculation of these losses, and the literature that is available generally focuses on gears moving through a lubricant rather than smooth surfaces. Early predictions of churning, perhaps because of limitation in the test equipment, vastly overestimated the torque lost at higher rotational speeds. Sample calculations on these earlier equations (not shown) determined that a modern, medium sized engine would struggle to rotate the discs of the CP-CVT as they rotated partially submerged through a traction lubricant, which is obviously unrealistic. It is only relatively recently that these equations were adapted and corrected to reflect modern experimental data.

2.6.5 Design Techniques

There are a vast number of different design approaches and methodologies available, for every aspect and stage of design. It is commonly agreed in a number of different literature sources that the design development procedure can be broadly divided into five stages: market, specification, concept design, detail design, and manufacture. Of most interest to this work are the concept and detail design stages, for which a number of tools exist, including Quality Function Deployment and in particular the use of a House of Quality. Both of these tools are designed to translate the voice of the customer into specific engineering requirements. Before they can be implemented however, the customer demands must first be established. In the

absence of a marketing survey, a number of existing literature sources were used to determine these customer demands, which include: low noise, reduced fuel consumption, ease of implementation, performance (power capacity), drivability (fast response), reliability/durability, and low capital cost. The aspects involved in a House of Quality have also been discussed, parts of which are implemented later.

In addition to design theories and techniques, a number of optimisation algorithms are discussed, all of which can be used for multi-criteria optimisation. A comparison of these techniques, (based on the author's opinion) is shown in Table 6.

Table 6: Comparison of optimisation technique

	Ease of Implementation	*Computational Demands*	*Accuracy of Solution*	*Premature Halting at False Maxima?*
Brute Force	Simple	High	High	No
Hill Climbing	Simple	Moderately Low	High	Yes
Iterative	Moderately Simple	Moderately Low	High	Possible
Genetic Algorithm	Complex	Moderate	Moderate	Unlikely
Swarm Optimisation	Moderate	Moderately Low	Moderately Low	Possible

CHAPTER 3: INITIAL DESIGN PROCESS: QUALITY FUNCTION DEPLOYMENT

Parts of this chapter have been presented by the author at the International Multi-Conference on Engineering and Technological Innovation, and have been accepted for publication in the International Journal of Mechatronics and Manufacturing Systems, Vol. 4, No. 1 (2011)

3.1 Introduction

3.1.1 Chapter Summary

This chapter uses one particular aspect of a process known as Quality Function Deployment (QFD) in order to determine the precise technical requirements of a transmission system intended for use in automotive applications. The specific requirements are used to further develop the CP-CVT design, the attributes of which address a specific market sector in an integrated approach. The discussion of the design attributes is centred on their interaction and effects on the final product.

The design approach used in this chapter incorporates a modified version of the House of Quality (HoQ) methodology, focusing on the characteristics of a mechanical transmission device influenced by a number of quality requirements. Customer demands are initially determined based on existing literature. These requirements are then translated into specific, measurable technical requirements. A relationship between the customer demands and the technical requirements is derived through the use of a modified HoQ, resulting in a numerical value for the relative importance of each technical requirement, highlighting those that should be optimised as a priority and those that are of lesser importance. Based on these technical requirements several design concept solutions are discussed, whilst the effect of changing the key component dimensions is also briefly analysed.

3.1.2 Overview of Quality Function Deployment

During the product development process, designers need to be able to respond to the rapid changes stimulated by technological innovations and changing customer demands. Product design incorporates many facets, such as requirement analysis, conceptual design, engineering design, product manufacturing, analysis of distribution and service support, and dispose/recycle aspects. As discussed by the two main schools of thought regarding the product development technologies: the scientific (Dixon, 1988) and the engineering, (Koen 1985), the design process has been optimised for different specific targets: design and assembly, design for disassembly, design for recycle, design for cost, design for reliability, design for serviceability, and design for environment. These design approaches focus on a single stage in the product life cycle and procedures for the integration of all these aspects have become an important aim for the engineering community.

The QFD method, which was initially proposed in Japan in 1966 (Akao, 1990) is a mechanism to translate the voice of the customer into engineering language. One particular aspect of the QFD process is the House of Quality (HoQ), which uses Pugh's concept selection method (Pugh, 1981). The House of Quality can be thought of as a conceptual diagram that provides the means for inter-functional planning between the customer demands and technical requirements (Houser and Clausing, 1988). This includes the desired attributes and the engineering characteristics together with their relative weights, with the aim of prioritising the characteristics by using the information stored in the HoQ. Typically a HoQ is produced using the following specific steps:

- Specify 'whats': determine important requirements by design experience to collect into the House of Quality;
- Translation of the requirements into 'hows', or technical characteristics;
- Construct the relationship matrix which will show the connection between 'whats' and 'hows';
- Analysis of the correlation matrix will show which requirements are more important by their weighting and how for a product each of the requirements are ranked;
- Based on the analysis, design targets are set and will be prioritised during the design process

During the design process the product design team must select which design features should be incorporated and furthermore which design features need to be improved upon. Traditionally, the complexity of the design process required the experience of the designer in order to make these selections. The House of Quality process attempts to remove some of this subjectivity by providing a continuous reference that shows directly how each design decision will affect customer satisfaction (Kim et al., 2000).

3.2 Methodology: House of Quality

3.2.1 Customer Demands

In order to determine the technical characteristics required from the CVT (the 'hows'), the first step is to determine what the customer (or in this case the driver) demands from an automotive transmission system ('whats'). The actual requirements of an automotive transmission system can be summarised as follows (Dutta-Roy, 2004):

- Disconnect and connect the engine drive train from the wheels as required.
- Reduce the rotational speed of the engine and increase torque output applied to the wheels
- Vary the transmission gear ratio as required by the driver to match the torque required at the wheels

These rather simplistic statements are the main functional requirements of an automotive transmission system and say very little about what the actual customer demands are. In order to determine this, a number of different sources have been examined. Hurst (1994) provides a good overview of typical factors that influence gearbox choice, including price, commercial life, installation convenience, maximum operating conditions (speed and torque), inertia, and noise, although these requirements are aimed more generally at gearbox selection rather than for automotive purposes. Similarly, Heilich and Shube (1983) make reference to power capacity, ratio range and a significant amount of detailed information regarding cost. Looking more specifically at CVTs for automotive purposes, Baudoin (1979) states that there are 6 key requirements that must be considered by any transmission designer: the ability to be mass produced at low costs; high efficiency; comfort (low noise and smooth ride); limited weight and volume; limited maintenance; and good reliability and life. Based on these literature sources, the following customer demands are proposed:

- Low noise
- Reduced fuel consumption/emissions
- Ease of implementation
- Performance (power capacity)
- Drivability (fast response)
- Reliability/Durability
- Low cost

The traditional House of Quality requires that each of these demands is assigned a value based on their perceived relative importance. In order to remove the subjectivity of assigning these ratings arbitrarily, it is proposed that a comparative matrix is used instead. This matrix contains

each of the demands as the column and row headers. Each demand is then given a score of 1, 3 or 9 relative to another where 3 indicates that the requirements are perceived to be of equal importance, 9 indicates that a row heading is perceived to be more important than a column heading, and 1 indicates the opposite. Using this system, Figure 3-1 shows the perceived relative importances of each demand for three different automotive applications: a typical family car (a), a performance or sports car (b), and a large passenger carrying vehicle (c).

Typical Family Car

	Low noise	Reduced fuel consumption	Ease of implementation	Performance	Drivability (fast response)	Reliability/Durability	Low cost	**Total**
Low noise		3	3	3	1	3	3	16
Reduced fuel consumption	3		9	9	3	9	3	36
Ease of implementation	3	1		3	3	3	3	16
Performance	3	1	3		3	3	1	14
Drivability (fast response)	9	3	3	3		3	3	24
Reliability/Durability	3	1	3	3	3		1	14
Low cost	3	3	3	9	3	9		30

(a) Typical Family Car

Performance or Sports Car

	Low noise	Reduced fuel consumption	Ease of implementation	Performance	Drivability (fast response)	Reliability/Durability	Low cost	**Total**
Low noise		1	3	1	1	3	3	12
Reduced fuel consumption	9		9	1	1	3	3	26
Ease of implementation	3	1		1	3	3	3	14
Performance	9	9	9		3	3	9	42
Drivability (fast response)	9	9	3	3		3	9	36
Reliability/Durability	3	3	3	3	3		3	18
Low cost	3	3	3	1	1	3		14

(b) Performance or Sports Car

Passenger Vehicle (bus or coach)

	Low noise	Reduced fuel consumption	Ease of implementation	Performance	Drivability (fast response)	Reliability/Durability	Low cost	**Total**
Low noise		3	9	3	9	3	3	30
Reduced fuel consumption	3		9	3	9	3	9	36
Ease of implementation	1	1		1	3	1	3	10
Performance	3	3	9		9	3	9	36
Drivability (fast response)	1	1	3	1		1	3	10
Reliability/Durability	3	3	9	3	9		9	36
Low cost	3	1	3	1	3	1		12

(c) Large passenger vehicle (bus or coach)

Figure 3-1: Customer demands relative importance matrices

The relative importance of each demand can be calculated by summing up each row. A summary of these is shown in Table 7.

Table 7: Customer demands relative important summary

Customer Demand	*Typical Family Car*	*Performance or Sport Car*	*Passenger Carrying Vehicle*
Low noise	16	12	30
Reduced fuel consumption	36	26	36
Ease of implementation	16	14	10
Performance	14	42	36
Drivability (fast response)	24	36	10
Reliability/Durability	14	18	36
Low cost	30	14	12

3.2.1.1 Low Noise

From Table 7 a reduction in noise is clearly of more importance to passenger carrying vehicles and to a lesser extent, family cars. In coaches and buses one of the most important aspects is passenger comfort; hence reducing engine noise is a large priority. Conversely for performance vehicles engine noise is less important and occasionally even desired. For automotive applications, the noise and vibration from the gearbox is typically much lower than the noise of the engine. This is especially true for traction drives, which have no metal on metal contact. In terms of the transmission system, noise reduction is thus achieved by allowing the engine to operate at lower speeds, ideally with only minimal changes in the engine's operating point. Important technical requirements for reducing noise thus include having a wide transmission ratio range, and minimal power variation.

3.2.1.2 Reduced Fuel Consumption

With the current global emphasis on reducing CO_2 emissions, it is perhaps not surprising that reduced fuel consumption is viewed as being of high importance. The improvements offered in fuel economy are often cited as the reason that CVTs are considered the future of automotive transmissions. This is perhaps more important for economical vehicles such as passenger or family cars rather than for performance cars, which inherently consume more fuel anyway. For a CVT to improve fuel consumption it must have a wide range of transmission ratios and be able to use this range effectively (low power variation). Furthermore, it must have a high transmission efficiency so additional energy losses within the CVT must be kept to a minimum.

3.2.1.3 Ease of Implementation

The automotive industry is generally regarded as being extremely slow to react to improvements in technology. Existing technologies such as a manual gearbox or suspension system have not changed significantly in the last several decades. Hence any new technology that is designed must be as easy to implement as possible. This is of more importance to family and sports cars, which are far more numerous and are more limited by the changes that consumers will accept. Larger vehicles, such as heavy goods vehicles or coaches/buses are generally more flexible, which is perhaps why ease of implementation is seen as being of least importance. Related technical requirements also include overall size, which must be at least comparable to existing transmission systems and to a lesser extent ratio range, which must be similar to existing gearboxes to ensure no significant additional engine modifications are required.

3.2.1.4 Performance

In this context, performance is considered to be the specific power capacity of the transmission. This is obviously of high importance for sports cars and larger vehicles, which have significantly higher torque demands; hence there is little point in using a CVT that has a rather limited torque capacity. The importance of this demand is highlighted by Heilich and Shube (1983), who list the typical horsepower ratings of traction drives of the time. These limits vary from only a few hp for certain designs, up to 100hp and beyond. It is important that modern CVT designs improve on this and are able to withstand a higher power per unit mass to be able to compete with current designs. Important characteristics are therefore torque capacity and mass, i.e. specific power capacity.

3.2.1.5 Drivability

Poor drivability is often claimed to be one of the reasons CVTs have not been widely implemented, despite the technology having existed for several decades. In early CVT designs that used friction rather than traction, slip would be so severe that the engine of the car would speed up significantly when more power was demanded until it eventually matched the speed of the drive belt simply because there was no grip between the two (Lang, 2000). Furthermore, one of the deficiencies in vehicles fitted with a CVT is the perceived lack of acceleration in comparison with standard transmission systems (Wicke et al., 2002). More recently a number of studies have looked at improving drivability (Wicke, Brace and Vaughan, 2000; and Smith et al., 2004), although it invariably involves using an arbitrary function that offers a higher engine speed for all throttle inputs (Smith et al., 2004), which is obviously not ideal. Drivability and response time is of more importance to performance and family cars, where drivers desire an immediate response, conversely in passenger carrying vehicles response time is often unavoidably poor simply because of the significantly higher vehicular mass. Without going into

detail regarding perceived acceleration and engine operating curves, the associated technical requirements of drivability are ratio range, power variation and to a lesser extent efficiency, since any losses will inherently reduce the response of the vehicle. Furthermore the component mass will have a significant effect on how quickly the CVT responds to changing demands, since a higher mass means a higher inertia.

3.2.1.6 Durability and Reliability

Durability and reliability are obviously important in any automotive application. Given the relatively high mileage of passenger vehicles, it is not surprising that reliability is seen as being more important. Furthermore, with a vehicle full of paying passengers, vehicle failure is obviously more costly. Both durability and reliability are difficult properties to quantify without extensive testing. Predictions of these factors for traction drive CVTs are usually based around experimental data, which provide design limits in terms of maximum contact pressure and traction coefficient (the ratio of tractive force to normal force), whilst operating close to, or at these limits increase the likelihood of failure. Since torque capacity is generally limited by these factors anyway, it is an important defining technical characteristic of durability. Additionally, the inherent efficiency of the device can affect durability/reliability. Larger losses increase the operating temperature of the lubricant, decreasing film thickness and thus increasing the probability of failure.

3.2.1.7 Cost

Heilich and Shube (1983) provide a detailed cost analysis of a traction drive including repayment periods, compound interest and depreciation however this goes far beyond the necessary information at this stage. A simple analysis involves the fuel savings that can be made, which are claimed to be around 10% (Boos and Mozer 1997), meaning a direct saving of 10% on fuel costs. This is somewhat offset by the additional cost of installing a CVT rather than a more popular transmission system, which can be quite significant due to the high precision and tolerance required. Cost is clearly of more concern to the users of standard family cars, rather than performance cars whom expect to pay a higher premium anyway. Likewise with passenger vehicles, the larger initial price of the vehicle means the additional cost of installing a CVT is somewhat absorbed. Ignoring the running costs of the vehicle, which are covered by reduced fuel consumption, the initial cost of the CVT can be simply reduced by reducing the amount of material required, hence mass and size are the defining technical characteristics.

3.2.2 Technical Requirements

Based on the customer demands discussed, the following technical characteristics ('hows') must be considered:

- Device efficiency
- Transmission ratio range
- Torque capacity
- Mass
- Length
- Diameter
- Output power variation

Traditionally, a House of Quality uses a symbolic matrix to relate the customer demands ('whats') to the technical characteristics ('hows'), with each symbol indicating the strength of the relationship. However, in order to determine the relative important of each technical characteristic it is more useful to use numeric values (1, 3 and 9, as before), with 1 indicating a weak relationship, 3 indicating a moderate relationship, and 9 indicating a strong relationship. This is shown in Figure 3-2.

		Application			Technical Characteristics						
		Family Car	Performance Car	Passenger Vehicle	Device efficiency	Transmission ratio range	Torque capacity	Mass	Length	Diameter	Output power variation
Customer Demands	Low noise	16	12	30	1	3					3
	Reduced fuel consumption	36	26	36	9	9		1			3
	Ease of implementation	16	14	10		3			9	9	
	Performance	14	42	36	3		9	9			1
	Drivability (fast response)	24	36	10	3	3		3			9
	Reliability/Durability	14	18	36	3		9				
	Low cost	30	14	12				3	3	3	
	Importance for Family Cars (Absolute)				496	492	252	324	234	234	386
	Importance for Family Cars (Relative)				0.21	0.2	0.1	0.13	0.1	0.1	0.16
	Importance for Performance Cars (Absolute)				534	420	540	554	168	168	480
	Importance for Performance Cars (Relative)				0.19	0.15	0.19	0.19	0.06	0.06	0.17
	Importance for Passenger Vehicles (Absolute)				600	474	648	426	126	126	324
	Importance for Passenger Vehicles (Relative)				0.22	0.17	0.24	0.16	0.05	0.05	0.12

Figure 3-2: Customer demands and technical characteristics relationship matrix

The advantages of using numbers to represent the relationship strengths are shown here. By taking the product of each relationship strength and its associated customer demand importance, each technical characteristic is automatically scored by application.

Although, as expected, there is some discrepancy between different applications, based on this matrix the most important considerations in transmission design are efficiency, ratio range and torque capacity.

3.3 Design Solutions

3.3.1 Efficiency

It can be reasonably assumed that the greatest losses within the traction device occur between the intermediary ball elements and the separator designed to force them to rotate and prevent adjacent elements from coming into contact (Cretu and Glovnea, 2005). Reducing the traction coefficient between the two surfaces would reduce these losses considerably. This could be achieved by having a deep-grooved separator, with the radius of curvature being only slightly larger than the radius of the ball elements. This would significantly increase the size of the contact between the separate and ball elements, decreasing the Hertzian pressure and also the traction coefficient. A larger reduction in separator losses could be achieved by ensuring a rolling contact. This is made more difficult however by the variation in the angle of the axis of rotation of the intermediary elements. A fixed-angle rolling separator would only provide a pure rolling contact for a small variation in the angle of the axis of rotation; at other times the rotation could be almost perpendicular to the separator, in which case the contact would be pure sliding. A variable angle separator would solve this but would incur significant additional costs and add an additional potential area of failure, reducing durability/reliability, making this a potentially unattractive solution.

Additional losses occur due to parasitic motions in the contact area (spin and creep) (Cretu and Glovnea, 2003 and Zhang et al., 2000) together with bearing losses and fluidic churning losses. A solution that reduced these losses without incurring significant penalties in other requirements would be perhaps more attractive. Optimising the CP-CVT in order to reduce these losses is discussed in detail in the following chapter.

3.3.2 Ratio Range

In order to compete with existing transmissions technologies, the CP-CVT must offer at least an equivalent range of ratios. Previous dimensions (Cretu and Glovnea, 2005) have offered a range of approximately 1:1.2 to 1:2.7, which is comparable, but not as good as most manual transmissions. The ratio range can be substantially improved by the addition of a planetary gear system. This system effectively deducts the engine speed from the transmission output speed, thus allowing a theoretical infinite selection of ratios up to a maximum limit. The total ratio range of the transmission can be increased to whatever value is required by altering the final-gear/differential ratio. The addition of a planetary gear system would also negate the need for an additional ‘clutch’ since there can be a net speed output of zero even when ‘in gear’, hence reducing the total weight of the transmission system.

An alternative solution would be to have an additional gear selection in series with the CVT. This could potentially double the range of ratios available at the cost of increased weight. In this system the CVT could operate in either a high or low regime, depending on the driving situation (urban or highway). The selection of this additional gear could be either automatic or manual. A manual option would restore some degree of control to the driver for certain situations such as driving downhill where the use of so-called 'engine braking' is preferable to the continuous use of regular brakes.

3.3.3 Torque Capacity

The torque capacity (and by extension, power capacity) of most traction drives is limited by the traction coefficient of the transmission fluid and the maximum allowable Hertzian pressure at the contact between adjacent parts. Previous research state that for a reasonable operational life the traction coefficient should not exceed 0.045 (Fuchs and Hasuda, 2004), meaning that for a normal load of 1000N at each contact, the tractive force should not exceed 45N. Assuming this value cannot be exceeded there are a limited number of changes that can be made to the CVT without increasing size and weight.

When certain criteria are met, traction drives have been shown to have the potential to outlast other key components in a vehicle. The durability of traction drives is generally limited by surface wear that occurs in normal lubrication conditions and by seizure due to scuffing, when for some reason, the lubricant film fails. To reduce the likelihood of this occurring, the traction coefficient must remain below its intended value. This can only be achieved by having a relatively large normal force, which leads to very high Hertzian contact pressures. One method to decrease this pressure is to increase the number of contact points by including additional intermediary elements. By increasing the number of contacts, the traction force per contact is lowered, reducing the normal force required at each contact. However, as shown in Equation 3.1, the benefit of increased contact points has diminishing returns.

$$p_0 = \left(\frac{1}{\pi}\right)\left(\frac{6W\overline{E}^2}{R_\sigma^2}\right)^{1/3} \tag{3.1}$$

Where:

- p_0 = Maximum contact pressure
- W = Normal contact force
- R_σ = Reduced radius of curvature of contacting surfaces $= \left((1/R_1) + (1/R_2)\right)^{-1}$
- $\overline{E}$ = Reduced elastic modulus $= \left\{\left(1-\nu_1^2\right)/E_1 + \left(1-\nu_2^2\right)/E_2\right\}^{-1}$

Hence whilst increasing the number of intermediary elements from four to eight would yield a reduction in Hertzian pressure of approximately 20%, increasing this number to twelve would only reduce pressure by an additional 12%. Equation 3.1 also indicates that a greater reduction could be achieved by increasing the geometric mean radius of the contacting elements, which can only be achieved by using larger components, increasing mass. Furthermore, larger components mean an increase in the rotational inertia of the CVT, which may also have a detrimental effect on the response time of the CVT, and hence may not be the best solution.

Typically CP-CVT component dimensions only allow a maximum of 3-4 ball elements and hence the only method of increasing the number of ball elements further is to incorporate two cages. Glovnea & Cretu (2006) proposed one such "double-cage design" in which the two cages are facing in opposite directions. In this design, the conical output disc is connected to a hollow cylinder, which has a gear train running around the circumference. This is connected to a separate output shaft, and is hence the input and output shaft are no longer co-aligned, as shown in Figure 3-3, decreasing the ease of implementation. Another disadvantage of this design is that a fixed gear ratio is automatically included, increasing the output speed and hence reducing the output torque. This would then require an additional correctional gear ratio after the transmission, adding an additional source of inefficiency.

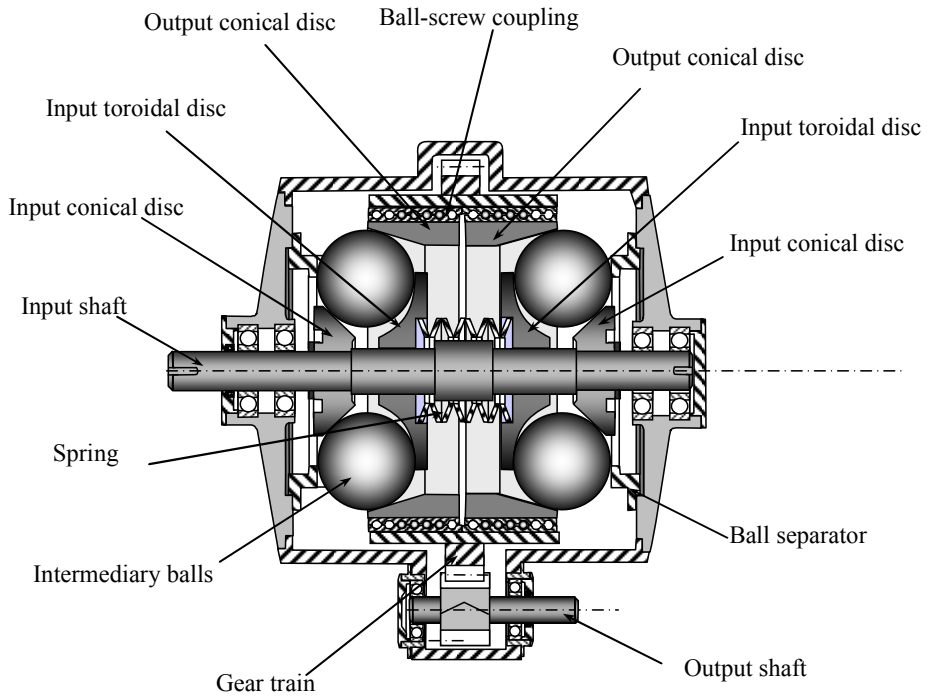


Figure 3-3: Double Cage Concept (Glovnea & Cretu, 2006)

One method of avoiding this would be to use two sets of discs facing in the same direction, as shown in Figure 3-4. To ensure that the two input discs maintain the same relative position, they would have to be connected via a connecting rod, which could pass through the second conical input disc, as shown in the figure.

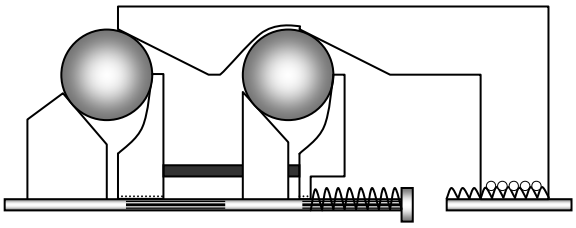


Figure 3-4: Alternative Double Cage CP-CVT

To stabilise and reduce vibration, thus reducing noise, a number of rolling elements could be placed between the exterior surface of conical output disc and the transmissions casing. Alternatively, the ball screw coupling could be repositioned and incorporated into this region, reducing the overall length of the design, as shown in Figure 3-5.

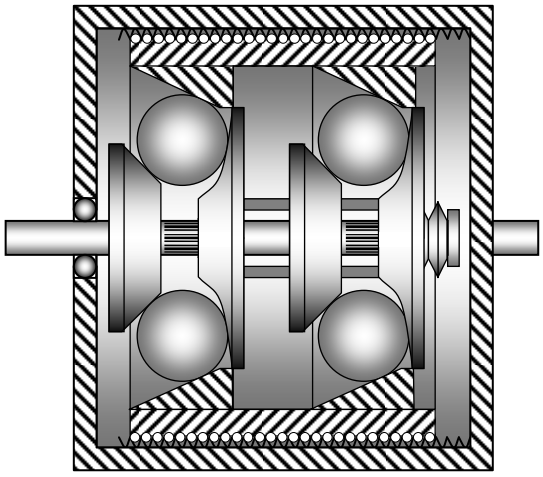


Figure 3-5: Double Cage Concept with external ball screw

It has been determined that the overall technical characteristics of a single or double cage design have an almost identical dependence on the fundamental component dimensions. Hence the geometric and dimensional optimisation shown in subsequent chapters can reasonable assume a single cage design, even if a two-cage design will be ultimately used.

3.4 Dimensional Discussion

Although later sections discuss dimensional optimisation in far more detail, it is useful to look briefly here at how and why certain dimensions affect the desired technical characteristics of the CVT. As discussed previously, the CVT primarily consists of a conical input disc, toroidal input disc, conical output disc and several ball elements. The parameter-defining dimensions can hence be described by 5 primary variables, as shown in Figure 2-9 (reproduced here for ease of reference)

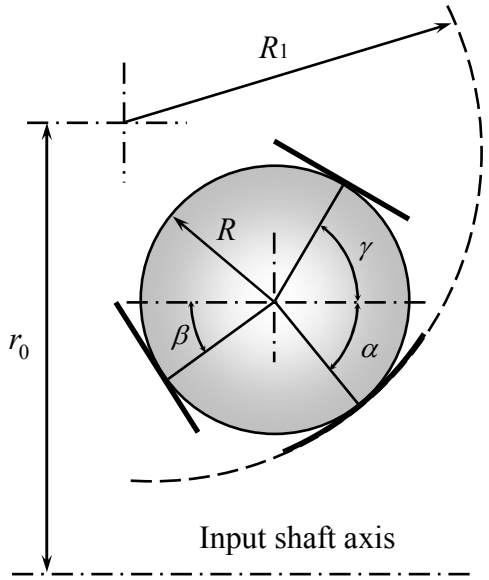


Figure 3-6: Explanation of dimensions of CP-CVT

- The contact angle of the conical input disc (β)
- The contact angle of the conical output disc (γ)
- The radius of the ball element (R)
- The concave radius of the toroidal disc (R_1)
- The secondary toroidal radius (r_0)

3.4.1 Conical disc angles

The contact angles of the conical discs are different from the other dimensions given that they do not directly affect the overall size of the CVT. Assuming all other dimensions can simply be scaled to adjust the CVT's parameters for various applications, the only independent dimensions that are unaffected by scale are the angles of the conical discs.

Since both the output conical disc and toroidal input discs each have one negative radius of curvature (concave curvature), it can be reasonably assumed that the highest Hertzian pressure is found at the conical input disc, even though the normal force has been shown to be higher at the conical output disc for any typical set of dimensions, as shown in Figure 3-7.

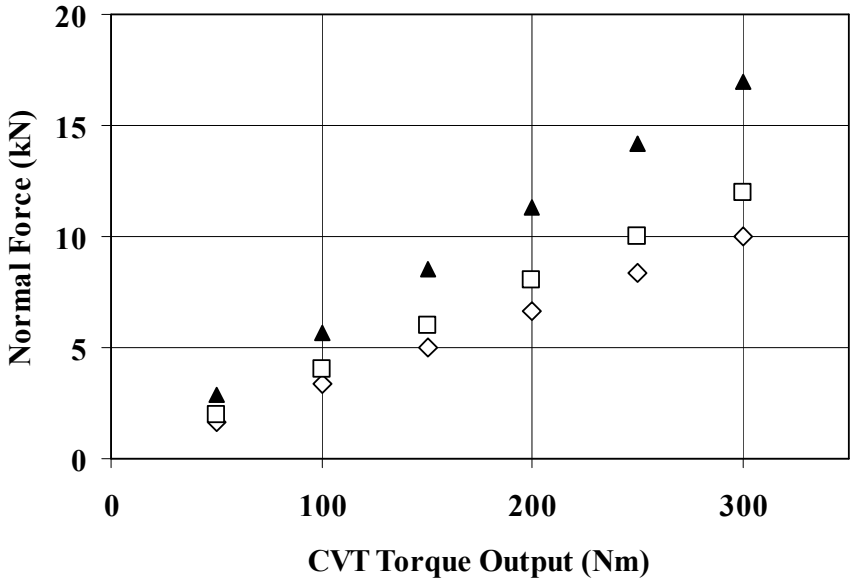


Figure 3-7: Effect of resistive torque on normal contact force per intermediary element (◊ = toroidal input disc; □ = conical input disc; ▲= conical output disc)

Given that the highest Hertzian pressure with the CP-CVT is found at the conical input disc, for a fixed torque output, the nature of the conical input disc has a large effect on the estimated peak Hertzian pressure. Hence for a fixed Hertzian pressure limit, decreasing the angle of the conical input disc ("flattening" the disc) significantly increases the radius of curvature, which in-turn enlarges the contact area, allowing a larger normal force, thus increasing the torque capacity. Furthermore, decreasing the angle of the conical input disc also forces the disc's contact with the ball element to move closer to the ball's axis of rotation, increasing the ratio range available. This would however require an excessively large normal force to transmit such a large shear force. One solution to this is to limit the range of movement of the toroidal disc, thus ensuring the conical input disc's contact with the ball is never too close to the ball's axis of rotation. Since the conical input disc has relatively small mass in comparison to other parts, irrespective of its dimensions, the overall mass and size of the CVT is not significantly affected by changing the angle. Based on this it is clear that better solutions should be found by having smaller angles of β.

Assuming the conical input disc has a relatively small angle ($\beta < 15°$) (the disc is relatively 'flat') then the ratio range available is potentially large enough that it can be somewhat ignored when determining other dimensions. However by limiting the range of motion of the ball elements for reasons discussed previously, the transmission ratio range is no longer sufficiently high to ignore. This range can be increased somewhat by increasing γ. Since the angle of the axis of rotation is independent of the γ, increasing its value simply allows the output disc to contact the ball element at a point further from the axis of rotation of the ball. This comes at a cost however, as increasing γ also significantly increases both the length of the conical disc, and by extension the overall length of the CVT as shown in Figure 3-8 (transmission ratio calculated using Equation 2.14). Both of these factors lead to a large increase of the overall mass, reducing the power-to-weight ratio. Provided the value of γ is less than approximately 75° this is no longer as much of an issue.

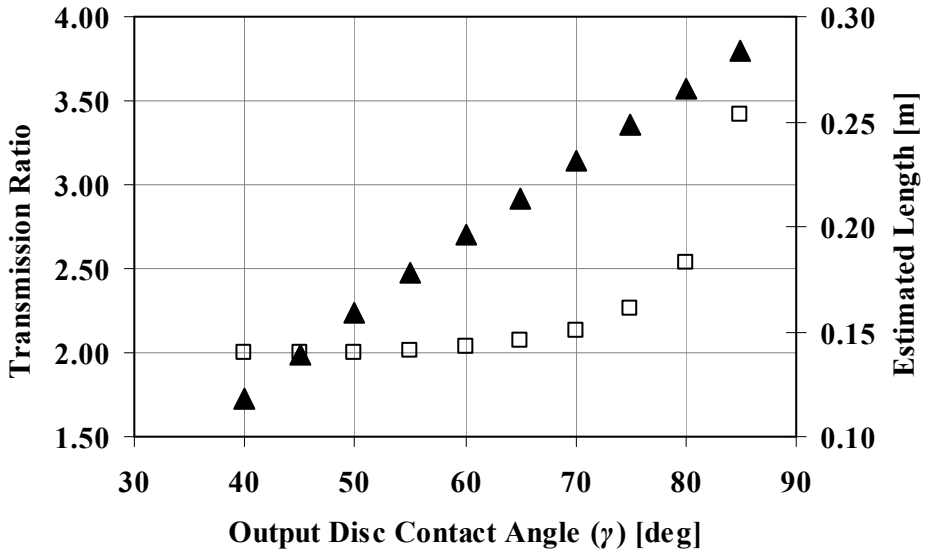


Figure 3-8: The effect of γ on max. transmission ratio (▲) and estimated length (□)

The nature of the ball-screw coupling has to be carefully designed to ensure that the axial force produced by it is sufficiently large to ensure an adequate normal force on the at the output contact. The traction fluid used in the device is designed to only operate up to a certain traction coefficient, beyond this, shock loading and temperature variations can cause the contact to breakdown, damaging the surfaces of the elements. Since there are two input contacts in most circumstances the traction coefficient is sufficiently lower at the inputs that it doesn't need to be considered. However with very large angles for both the conical input and output discs, this is no longer the case. Since, as shown, there are more obvious benefits to having a smaller conical input disc angle; it is hence the conical output disc angle that must be limited. For example,

with a β value of 10°, γ must be less than approximately 50°, which conveniently still allows a considerable reduction in Hertzian pressure and mass, whilst maintaining a large enough range of transmission ratios.

3.4.1.1 Toroidal radii

The shape of the toroidal disc can be described by two radii; R_1 and r_0, as shown in Figure 3-6. Based on this figure, the limits of the radii are fairly straight forward; the radius of the curvature of the toroidal disc must clearly be greater than the radius of the ball. Furthermore if it is too large there will be less room between each ball element when they are closest to the shaft. This will either reduce the number of ball elements that will fit, or reduce the amount of space available for the separating cage. Larger values of R_1 will also force each ball element to travel a greater distance when changing ratio, worsening response time, whilst values too close to the radius of the ball elements will increase the accuracy required from the loading system to achieve a specific transmission ratio.

The space between adjacent ball elements can be increased by using a larger value of r_0. This will enlarge the overall diameter of the CVT however, increasing mass and reducing the power-to-weight ratio. Larger values of both toroidal radii allow the use of a larger ball elements, which will in-turn increase the contact diameters, reducing Hertzian pressure, and thus increasing torque capacity. Using Equation 2.14, the greatest range of transmission ratios can be accomplished by reducing the value of r_0 and increasing the value of R_1, as shown in Figure 3-9.

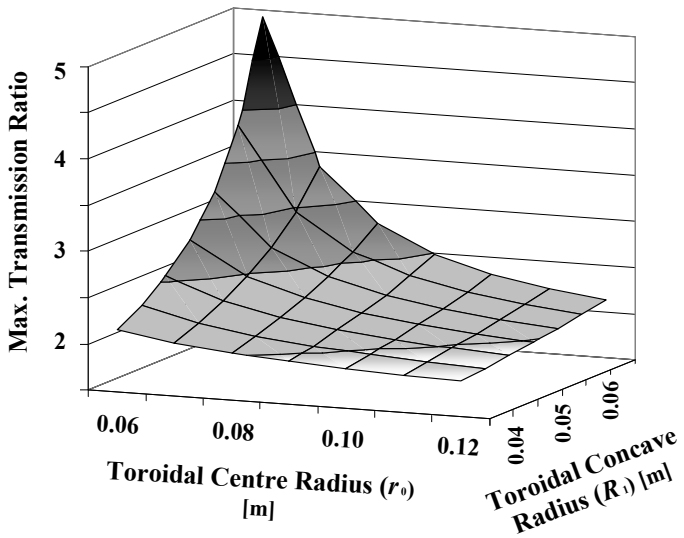


Figure 3-9: Effect of toroidal dimensions on maximum transmission ratio

3.4.1.2 Radius of ball elements

The parameter most affected by the radius of the ball elements (R) is the Hertzian pressure. Whilst having smaller values of R enable the use of more ball elements, decreasing the normal force per ball, a larger decrease in Hertzian pressure can be achieved by having fewer, but larger balls, as indicated in Equation 3.1. This would however significantly increase the mass of each ball, and therefore the overall mass of the CVT.

Experimentation with combinations of dimensions using Equation 2.14 indicates that ratio range is not significantly affected by the radius of the ball elements, however roughly a 10% increase can be achieved by having the ball radius as close to the toroidal radius as possible.

Perhaps counter-intuitively, the overall length of the CVT actually decreases for large values of R. The reason for this is that as the ratio of R / R_1 increases, less movement is required from each ball element, reducing the movement required from the conical output disc, which is the primary parameter that dictates the overall length of the CVT.

Using rudimentary calculations, the overall mass of the CVT was found to increase significantly with ball element radius (more accurate mass calculations are presented in later chapters). This effect is obviously more pronounced as the number of ball elements increases, as shown in Figure 3-10, highlighting the significant contribution the ball elements have to the overall mass of the CVT components.

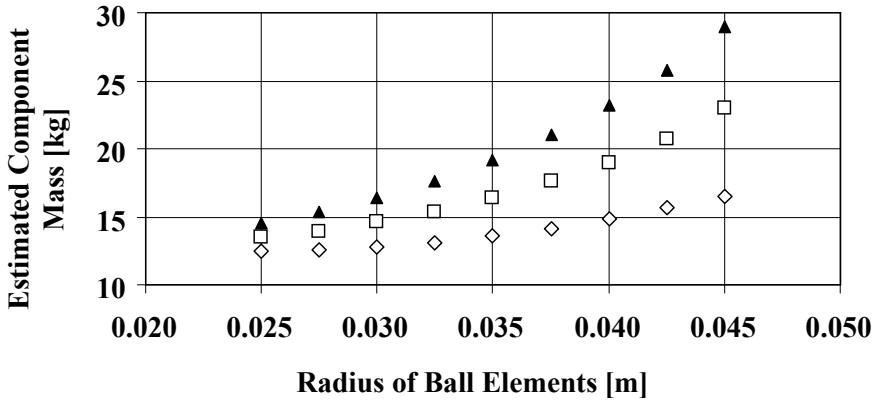


Figure 3-10: Effect of ball radius on mass (◊ = 4; □ = 6; ▲= 8 ball elements)

3.4.2 Summary

Looking at each of the component dimensions individually, a number of preferences have been derived regarding the nature of each dimension. These are summarised in Table 8. More advanced detailed calculations regarding the technical characteristics are presented in Chapter 5. However, since later optimisation techniques use advanced algorithms to determine each dimension, it is useful at this stage to have some indication about the typical dimensions that should be generated.

Table 8: Summary of preferred component dimensions

Parameter	*Conical input angle (β)*	*Conical output angle (γ)*	*Toroidal disc concave radius*	*Secondary Toroidal Radius*	*Radius of ball elements*
Maximum Torque Capacity	Smaller	Smaller	Unaffected	Unaffected	Larger
Lowest mass	Unaffected	Smaller*	Smaller	Smaller	Smaller
Largest ratio range	Smaller	Larger	Larger	Smaller	Unaffected
Smallest size	Unaffected	Smaller*	Smaller	Smaller	Larger

* Near-constant when less than approximately 70°

Given that both mass and Hertzian pressure have an influence on two of the most important design criteria, these are the two measurable parameters that should be optimised as a priority, even though ratio-range is seen as a more desirable attribute overall. To achieve this it is clear both angles should be relatively large, whilst both toroidal radii should be small as possible. The one dimension that requires a compromise is the radius of the ball elements. If each of the other dimensions has already been established than the radius of the ball elements should simply be large enough to reduce the Hertzian pressure to an acceptable value. If they are larger than this than both the power-weight ratio and the response time would suffer as a result.

3.5 Conclusions

This chapter has presented an application of one aspect of the Quality Function Deployment (QFD); the House of Quality. This methodology has been applied to the design of a CVT system, with the aim of improving the design in terms of certain measurable technical requirements. These requirements, which were derived from customer demands, were rated according to their perceived influence on overall customer satisfaction for a variety of specified applications.

Once the technical requirements were determined, they were then analysed and improvements were suggested individually based on the priorities generated by the HoQ. Further analysis indicated that the interrelation of the requirements necessitates that all must be considered at the same time to ensure that improving one area does not significantly worsen another. Hence in order to optimise the design further, physical properties such as mass, ratio range, and efficiency need to be calculated and quantified simultaneously with the aim of meeting specific measurable targets.

Certain design solutions such as variable separator geometry or the use of a planetary gear system have been discussed; however the most significant improvement in the technical characteristics of the CVT can be achieved through optimising the component dimensions. The following sections look at achieving this. Firstly by looking at efficiency in isolation (Chapter 4), since this is by far the most complex characteristic to predict and quantify, then subsequently by looking at all of the listed technical characteristics together (Chapter 5).

CHAPTER 4: EFFICIENCY OPTIMISATION

Parts of this chapter have been presented by the author at the Leeds-Lyon Symposium on Tribology, 2010, and have been accepted for publication in Proc. IMechE Part J: Journal of Engineering Tribology (awaiting publication details)

4.1 Introduction

4.1.1 Chapter Summary

It has been established in the previous chapter that efficiency is one of the critical technical characteristics required from a transmission system. Hence in order to compete and be viewed as a reasonable alternative to other transmission systems (including alternative CVT designs) the overall efficiency needs to be improved. To achieve this improvement, this chapter looks at the sources of loss within the CP-CVT to determine the best set of component dimensions that will provide the lowest losses. The methods used for calculating the losses are based on numerous sources from existing literature, whilst certain simplifications are proposed in order to make the calculations manageable and repeatable. Detailed example calculations are shown for one particular set of dimensions based on those found in previous literature regarding the CP-CVT. These calculations highlight the complexity of predicting efficiency, and show possible methods of simplifying certain calculations by making reasonable assumptions and approximations.

To determine the theoretically highest transmission efficiency, a novel optimisation technique is proposed and developed based on existing particle swarm and ant colony techniques. Rather than attempting to determine a single set of dimensions, the method is capable of quickly producing approximate values and determining regions of interest within the search space, significantly reducing computational time. In addition to dimensional optimisation, further discussion is also made regarding design modifications that can reduce these losses. The search algorithm is also compared to alternative methods of optimisation, and the effectiveness of the algorithm used is discussed.

4.1.2 Sources of Inefficiency

Across the CVT there are several sources of losses that contribute to the overall efficiency of the device. The primary sources of losses come from the elastohydrodynamic contact (creep and spin), churning of the traction fluid, bearing losses, and losses due to the ball separator (cage). Each of these losses, which are described in more detail in the following sections, will contribute to a reduction in the torque or rotational speed of each element, and hence for accuracy losses must be calculated progressively, starting from the input shaft. The power available at the output shaft can then be compared to the initial power available in order to determine the overall efficiency of the device, as shown in Figure 4-1.

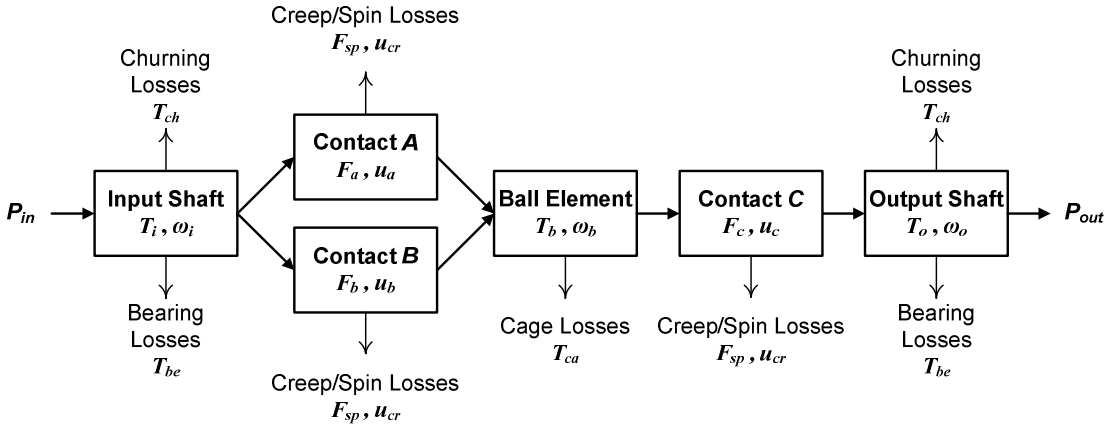


Figure 4-1: Power loss flow diagram

In order to better understand the concept of contact and cage losses, Figure 4-2 (based on Zhang, Zhang and Tobler, 2000) shows the direction of the tractive and spin forces on the contact between a single ball element and an input disc, together the contact between the cage and ball.

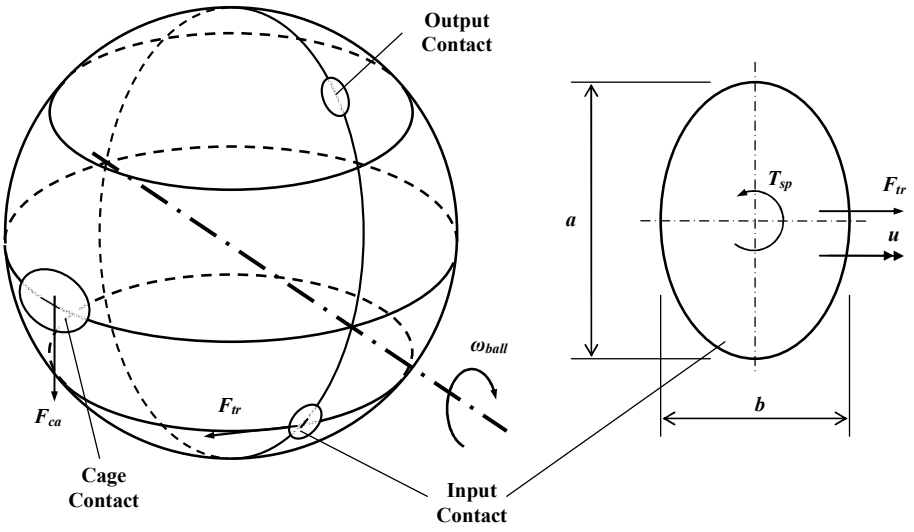


Figure 4-2: Directions of spin and traction on single ball element

The tractive force applied from the input disc acts to rotate the ball element about its own axis. Opposing this force is the traction between the cage and the ball element, which acts in the opposite direction to the rotational motion of the ball element. Within the contact between the ball element and the discs there also exists a 'spin' torque, which partially opposes the tractive force and arises because the axis of rotation of the ball elements and the input/output shafts are not parallel. Note that for this particular CVT design there is an absence of side slip, since the axes of rotation of all elements are co-planar.

4.1.2.1 Creep

In toroidal traction drives there is no direct contact between the intermediary element and the discs. Force is instead transmitted through a special Elasto-Hydrodynamic (EHD) traction fluid. When this fluid is put under extreme pressure it essentially takes on the properties of a solid. This so-called 'elastic–plastic transition' pressure is approximately 0.69 GPa for a widely used, current traction fluid (Ohno, 2006). An EHD fluid's properties are described by a traction fluid coefficient (μ), which for current fluids has maximum value of 0.12 (Evans & Johnson, 1986). However traction fluids do not transfer force perfectly and hence some slip will still occur between the two surfaces, known as creep. Creep is therefore defined as the tangential velocity difference between contacting elements which arises from the small amount of slip that is required in the rolling direction in order to generate traction (Tevaarwerk and Johnson, 1979).

Creep losses are typically relatively small since traction drives are designed to operate in the linear region of the traction coefficient curve to avoid failure of the lubricant film due to overheating. In order to determine the magnitude of the slip, Figure 4-3 shows typical traction curves for a high-traction fluid, Santotrac50 (Anghel, Glovnea and Spikes, 2004), where S is defined as in Equation 4.1.

$$S = \frac{2|u_1 - u_2|}{u_1 + u_2} \tag{4.1}$$

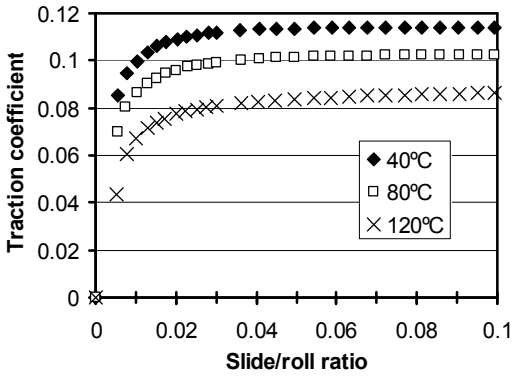


Figure 4-3: Traction coefficient as a function of slide-roll ratio

It has already been established to ensure a reasonable operating life and thus acceptable level of wear, the maximum traction coefficient should not exceed 0.0045 (Fuchs and Hasuda, 2004). From Figure 4-3 it is clear that if this maximum traction coefficient value is imposed then the slide to roll ratio will be less than 0.01, hence creep losses will be relatively small. In order to simulate this behaviour, a simple relationship was determined between the slide-roll ratio and traction coefficient based on a polynomial curve fit to the experimental data, although for the majority of realistic traction coefficients this is unnecessary since the curve is a straight line below a traction coefficient of approximately 0.06. This allows the slide-roll ratio to be calculated for any contact if the normal and traction forces are known. The slide-roll ratio can then be used to determine the magnitude of the creep losses.

The creep loss in terms of velocity (u_{cr}) can be defined as the difference between the linear speeds of the surfaces at either side of the contact, i.e.

$$u_{cr} = u_1 - u_2 \tag{4.2}$$

At any instance the value of u_1 will be known, whilst u_2 can be calculated from the slide-roll ratio. Rearranging Equation 4.1 in terms of u_2:

$$u_2 = u_1 \frac{(2 - S)}{(2 + S)} \tag{4.3}$$

Combining Equations 4.2 and 4.3:

$$u_{cr} = u_1 \left[1 - \frac{(2 - S)}{(2 + S)} \right] \tag{4.4}$$

Hence the creep velocity can be calculated using only the tangential velocity at the input side of the contact, together with the slide-roll ratio.

4.1.2.2 Spin

Spin losses arise from the angular velocity differences between two contacting surfaces in the *z*-axis of the contact. The effect of spin on traction has been of interest to researchers for a number of years due to its negative effect on the efficiency of traction drives and bearings (Newall and Lee, 2003). In general terms, the angular velocity of spin can be defined as shown in Equation 4.5.

$$\omega_{sp} = \omega_{1z} - \omega_{2z} \tag{4.5}$$

Figure 4-4 shows an arbitrary contact between two, non-parallel axes. From this Figure, and Tanaka (1989), it can be shown that:

$$\omega_{1z} = \omega_1 \sin\theta_1 \tag{4.6}$$

$$\omega_{2z} = \omega_2 \sin\theta_2 \tag{4.7}$$

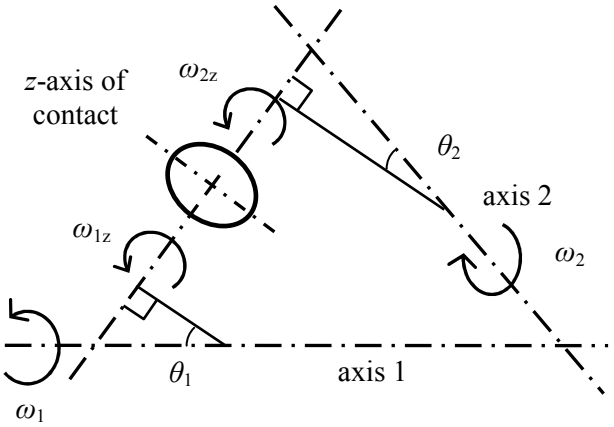


Figure 4-4: Spin velocity as a function of axis geometry

Research typically focuses on how spin will affect the traction curves for a given contact, however for the purposes of this study; it is of more interest to determine the total power lost to spin, from which the tractive force lost to spin (F_{sp}), (which acts in the opposite direction to the tractive force) can be calculated. Johnson (1985) states that the twisting moment that arises about the spin axis (*z*-axis of contact) due to traction is given as:

$$T_{sp} = \frac{32(2-\nu)}{9(3-2\nu)} G b^3 \psi \tag{4.8}$$

Where ν is the Poisson's ratio of the material and b is the contact semi-width (shown in Figure 4-2), which can be calculated from Hertzian theory and ψ is known as the dimensionless spin factor, which is defined as:

$$\psi = \frac{2\omega_{sp} c}{(\omega_1 r_1 + \omega_2 r_2)} \tag{4.9}$$

For simplicity we can take G as the average fluid shear modulus, which according to Hirst and Moore, (1978) has been shown to be:

$$G = (3/8)\pi m (h_c / b) P_m \tag{4.10}$$

In this equation h_c is the central film thickness, which according to Chittenden et al. (1985) can be calculated to from:

$$\frac{h_c}{R_\sigma} = 3.0\left(\frac{2U\eta}{\overline{E}R_\sigma}\right)^{0.67}\left(\alpha\overline{E}\right)^{0.49}\left(\frac{2W}{\overline{E}R_\sigma^2}\right)^{-0.073} \tag{4.11}$$

where U is the entrainment speed (average of u_1 *and* u_2), W the applied load, $\overline{E}$ the reduced Young's modulus, R_σ the reduced radius in the rolling direction, η the lubricant dynamic viscosity and α the lubricant pressure viscosity coefficient, both of the latter at the contact inlet temperature.

By multiplying the twisting moment that resists spin by the spin velocity, we can obtain an expression for the power lost due to spin in each contact:

$$P_{sp} = \frac{32(2-\nu)}{9(3-2\nu)}Gb^3\psi\omega_{sp} \tag{4.12}$$

The tractive force that is lost in the direction of traction can now be determined from the power lost and the linear tangential velocity of the element in question (u_1), as shown in Equation 4.13.

$$F_{sp} = \frac{P_{sp}}{u_1} = \frac{32(2-\nu)}{9u_1(3-2\nu)}Gb^3\psi\omega_{sp} \tag{4.13}$$

4.1.2.3 Cage Losses

Cage losses are unavoidable losses that occur due to the contact between the ball separator and the ball elements. Whilst there have been patented solutions with modified spherical elements which are supported on shafts rotating on rolling element bearings that reduce cage losses, they also considerably increase the complexity of the device and implicitly the cost. If a simple ball separator is used similar to that typically found in a rolling bearing, the nature of the EHD contact will be pure sliding. The surface geometry of the separator can be specifically designed to increase the contact area and hence reduce the Hertzian pressure, reducing the traction coefficient. Although it could possibly be higher, a μ value of 0.02 will be used for calculations, with the knowledge that the calculated transmission efficiency is optimistic. Comparatively between different solutions, this will have little influence. Knowing this, torque lost by the spherical elements to the cage can be defined as:

$$T_{ca} = \mu WR \tag{4.14}$$

where W is the normal force acting on the contact between the cage and the elements. This can be calculated from the input torque and the immediate distance between the centre of the ball element and the centre of the input shaft (r), as shown in Equation 4.15.

$$W = \frac{T_{in}}{r} \tag{4.15}$$

Combining Equations 4.14 and 4.15:

$$T_{ca} = \frac{RT_{in}\mu}{r} \tag{4.16}$$

4.1.2.4 Bearing Losses

Both the input shaft and output shaft are subjected to a very high axial load together with a smaller radial load; hence the ideal bearing for this application would be a tapered roller bearing. Since the spherical elements are designed to be distributed equally around the circumference of the input discs, the only radial loads acting on the input and output bearings would be due to the mass of the elements attached to each respective shaft. These radial loads are insignificant compared to the axial loads acting on both the input and output bearings, which will be equal to the forces produced by the ball-screw coupling together with any pre-compression applied to the spring acting on the toroidal disc.

Witte (1973) determined from empirical data that the frictional torque of tapered roller bearings due to axial loads can be calculated using Equation 4.17 (corrected for change in units from original source to give output in Nm).

$$T_{be} = 3.84 \times 10^{-5} k F_{axial}^{1/3} \sqrt{(\omega \nu)} \tag{4.17}$$

Where v is the viscosity of the lubricating oil and k is a factor based on the geometry of the bearing used. For a typical tapered roller bearing of an appropriate size for this application, k was found to be approximately 10.

4.1.2.5 Churning Losses

Churning losses arise due to the viscous dissipation that occurs as a disc rotates through a liquid. This has been studied empirically for a long time, with no overall agreement on the magnitude of the losses. Generally studies have focused on the losses that occur due to a spur gear rotating in oil; however one of the few studies that looked at a smooth disc was conducted by Boness (1989), who determined that the losses due to churning in terms of torque can be calculated as:

$$T_{ch} = \tfrac{1}{2} C_M \rho \omega^2 R_{im}^{\ 3} A_{im} \tag{4.18}$$

Where R_{im} and A_{im} are the submerged surface radius and area respectively and C_M is the moment coefficient. Luke and Olver (1999) found that Boness' original equations vastly overestimated the torque lost due to churning at higher rotational speeds and hence derived their own equation based on experimental results for C_M, which for turbulent flow can be calculated using Equation 4.19

$$C_M = \frac{5.34 \times 10^4}{\mathrm{Re}^{1.379}} \tag{4.19}$$

Where 'Re', according to Luke and Olver (1999) is the Reynolds number of the partially submerged disc's interaction with air, which in this instance can be calculated using Equation 4.20.

$$\mathrm{Re} = \frac{\omega R_{im} L_{im}}{\nu_{air}} \tag{4.20}$$

4.2 Methodology

4.2.1 Example Calculation

To demonstrate the equations shown, a full set of example calculations are shown in Appendix A.1. These calculations use a typical set of component dimensions as given by Cretu and Glovnea (2005), which are shown in Table 9 and Figure 4-5. In order to calculate losses an assumption needs to be made about the power supplied from the engine. Assuming a typical power output of 62.8kW at 5000rpm, this gives a torque output from the engine of 120Nm. The CP-CVT is assumed to have four intermediary elements, whilst the traction fluid is assumed to be a constant 80°C.

Table 9: Typical set of existing CP-CVT component dimensions

Parameter	*Value*
β	45°
γ	75°
R	0.06m
R_1	0.09m
r_0	0.12m

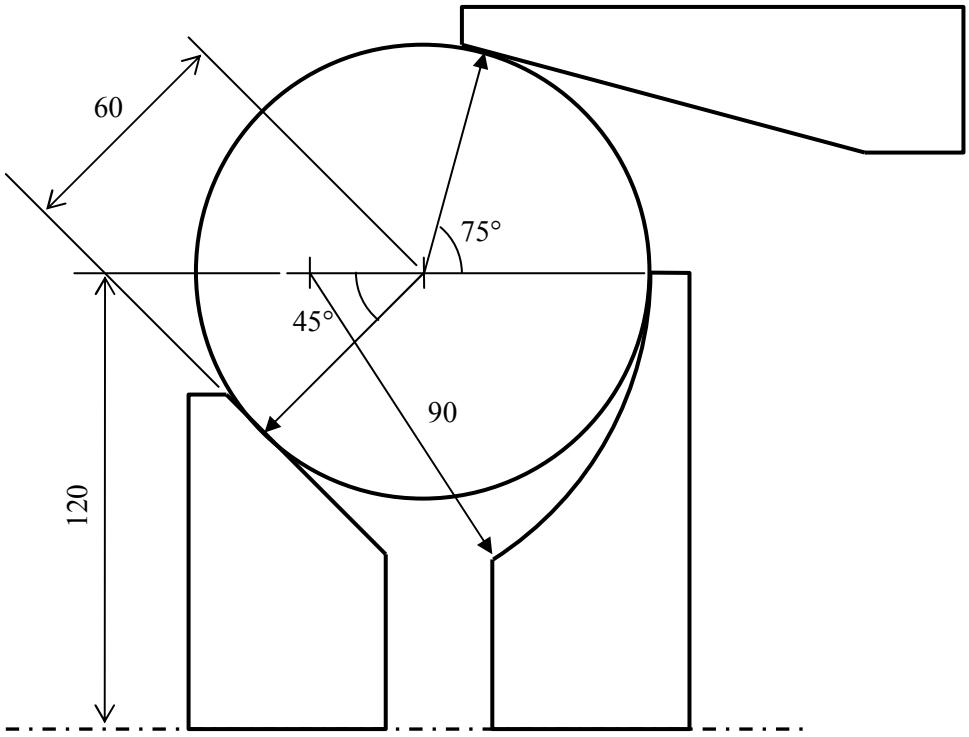


Figure 4-5: Scale sketch of CVT dimensions shown in Table 9

The CVT has a wide range of operating conditions depending on the transmission ratio *i*, and hence the contact angle between the ball elements and the toroidal disc (α in Figure 2-9), which for convenience and simplicity can typically vary between 0 and 1 radians (0-57°). The exact running efficiency of the CVT will depend on how much time the CVT spends at each ratio, however test calculations on a wide variety of dimensional configurations indicated that an α value of 17° generally gave the best indication of the average efficiency of the CVT, significantly reducing the number of calculations that would be required to determine the efficiency of the CVT at every operating point.

A summary of the calculated losses is shown in Table 10, Figure 4-6 and Figure 4-7. This shows that the total losses for these particular dimensions are 4.83kW when the power supplied to the input is 62.8kW. This gives an overall efficiency of 92.3%. This is not particularly high, largely because the dimensions are far bigger than they need to be for the power applied to the input. This is highlighted by the particularly low Hertzian pressure shown in Appendix A.1, which indicates that the dimensions could be significantly smaller, decreasing certain losses, without adversely affecting the overall performance and life. Further calculations indicate that these dimensions would allow more than 180Nm of torque at the input without exceeding the Hertzian pressure limit, an increase of 50%.

Table 10: Summary of calculated losses

Loss Source	*Loss (W)*	*Contribution to Total Losses*
Input shaft bearing	319	6.6%
Churning of input discs	1094	22.6%
Creep at input contacts	146	3.0%
Spin at input contacts	79	1.6%
Cage	1742	36.0%
Creep at output contact	133	2.7%
Spin at output contact	45	0.9%
Churning of output disc	1053	21.8%
Output shaft bearing	216	4.4%
TOTAL	**4827**	

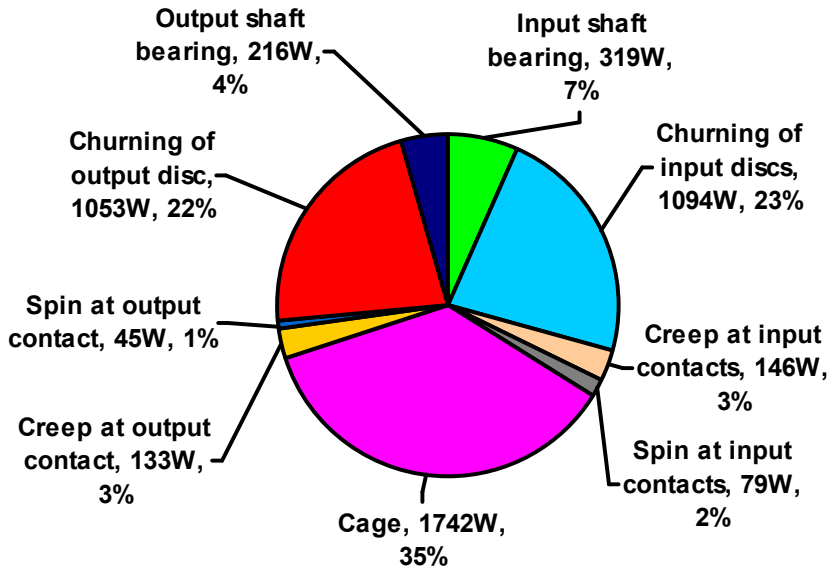


Figure 4-6: Contributions of each loss source for a typical set of CVT dimensions

Figure 4-7 shows a summary of the power flow and losses for this example in quasi-Bond graph notation.

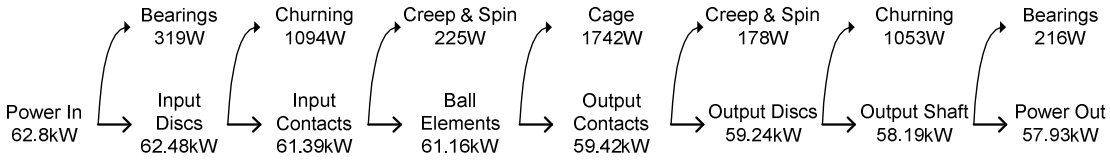


Figure 4-7: Calculated loss power flow

4.2.2 Optimisation Algorithm

The purpose of this dimensional optimisation problem is to determine the 'best' combination of dimensions in order to satisfy specific requirements. This problem can be viewed as a search problem, in which there is a *5*-dimensional search-space of possible solutions, with each dimension representing one of the five key component dimensions that can fully describe the CP-CVT (β, γ, R, R_1 and r_0). Each solution within the search space can be scored depending on its ability to meet the desired target, which in this case is efficiency.

For this type of problem there are a number of possible search techniques that could be employed, including brute force, hill-climbing, or swarm-optimisation.

4.2.2.1 Alternatives: Brute force search

A brute force search is the simplest and most computationally demanding search optimisation. Every possible solution is evaluated in order to determine the best solution. This is not practical in any problem in which there exists a large number of potential solutions. The precise number of solutions that have to be evaluated is determined by the desired resolution of the dimensions. With five input dimensions, even a relatively low resolution requires a significant number of calculations, and is not a particularly economical method of optimisation.

An alternative approach involves iterative searches, which initially uses large intervals to broadly assess the search space and then focuses on the area that yields the best results. This can often be an attractive approach however this assumes that the nature of the search space is relatively simply, without a large number of local, false-maxima. The problem with this approach is that if the search initially follows the wrong path, the best solution will never be found.

4.2.2.2 Alternatives: Hill Climbing

In its simplest form, hill climbing involves picking a random starting point, then altering each input dimension by a small interval to determine if there is a better adjacent solution, this process is then repeated until there are no better solutions surrounding the current one. Although there are variations of this, the methods generally suffer the same flaw of premature halting in which the algorithm could stop if there is no better adjacent solution. This can often occur if there are local false-maxima in the search space which are not the true maximum point.

An extension of this is random-restart hill climbing, which progressively uses multiple starting points, with better solutions replacing earlier stored solutions. Although this somewhat reduces

the problem of false solutions, it still requires a relatively simple search space in order to continue successfully.

4.2.2.3 Particle Swarm Optimisation

If hill-climbing is considered a particle search method, where the particle moves around the search space according to the adjacent solutions, Particle Swarm Optimisation (PSO) could be seen as an extension of this (Kennedy and Eberhart, 1995). In PSO a number of particles are used simultaneously whilst their movement around the search space is governed not only by local/adjacent solutions, but also by information learned from other particles. If run ad infinitum, all of the particles would tend to swarm to a particular, high-scoring solution, however this is never guaranteed.

Although far less computationally demanding than a brute-force search, generic particle-swarm techniques can still be inefficient methods of optimisation. Each movement through the search space still requires the evaluation of every adjacent solution. With five dimensions this requires ten full evaluations (or 26 if diagonal movement is permitted) per step. Furthermore, the solutions evaluated will be generally focused around a relatively small region, and hence these methods offer no additional information about the remainder of the search space.

4.2.2.4 Fuzzy Swarm Optimisation

Given the disadvantages of these standard optimisation techniques, a new form of algorithm is proposed based on a combination of Ant Colony Optimisation (Dorigo, Birattari, and Stutzle, 2006) and Particle Swarm Optimisation (Kennedy and Eberhart, 1995). This simple algorithm is capable of not only quickly finding good solutions, but also providing a significant amount of additional information regarding the nature and sensitivity of each of the input parameters. The algorithm initially involves distributing a number of particles throughout the search space at random in order to obtain a general idea of the nature of the search space, much like ants initially wander randomly from their home to find food. Further particles are then released, with a higher probability of falling closer to previous solutions that have demonstrated a higher fitness score (higher efficiency). Hence in the same way that ants tend to follow the paths created by previous ant trails, particle will tend to converge on a number of particular points. Each of the particles does not ‘move’ around the search space, thus reducing the need for continuous adjacent solution checking. As the algorithm runs, its knowledge of the search space increases and it is able to assign particles to the areas of interest with greater accuracy. In order to test the effectiveness of a number of different search algorithms, a 2-dimensional array of values was randomly created, with values ranging from 0 to 1 which together form a contoured surface. This is designed to represent a typical complicated function, with two inputs, with the

aim of finding the largest value, and the location at which it occurs. An example of a contoured surface created is shown in Figure 4-8. The method of creation of these surfaces is shown in Appendix A.2.

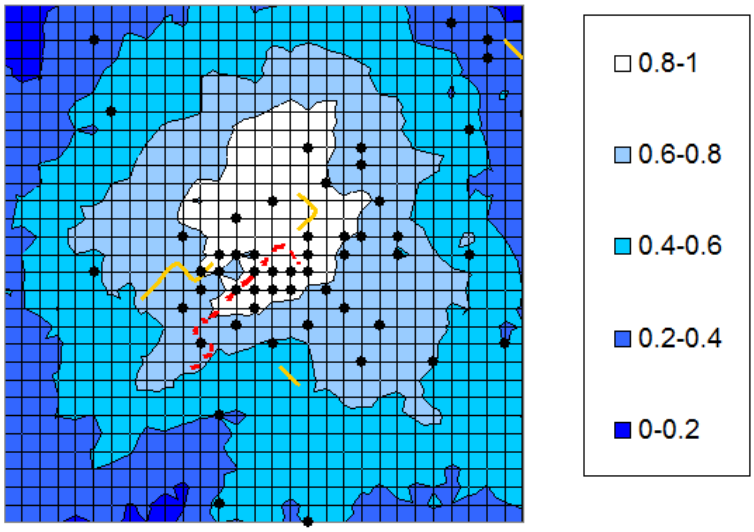


Figure 4-8: Demonstration of fuzzy-swarm algorithm operation (simple terrain)

The aim of this application of the fuzzy-swarm algorithm was to determine the highest peak of the surface. In order to achieve this, the following steps were taken:

1. Initially 10 points were assigned randomly and the height determined for each point.
2. 25 new points were then created with a 50% probability of being at a completely new, random location (P_{new} = 50%) and a 50% probability of falling close to a previous, high scoring point.
3. A further 25 points were then created, with a 10% probability of falling at a completely new location (P_{new} = 10%), and a 90% probability of falling close to an existing point.

If it is decided that a new point should fall close to an existing point, the existing point is chosen using a roulette style selection (explained in more detail later), with higher scoring points having a higher probability of attraction. The exact location of the new point is then determined using a spread factor, which forces the new point to stay within a pre-determined radius of the existing point. The spread factor is hence defined as the percentage deviation from the selected existing point.

Figure 4-8 also shows how hill-climbing (red-line) and random-restart hill-climbing (yellow lines) traverse the search space, whilst each black circle represents a point selected by the fuzzy-swarm algorithm. This figure shows that both hill-climbing and the fuzzy-swarm algorithms were able to find the regions containing the highest points, which they were able to do for the

majority of randomly-generated terrains with a single peak. However, when the terrain is more complicated (Figure 4-9); the hill-climbing techniques become hopelessly lost in false maximum.

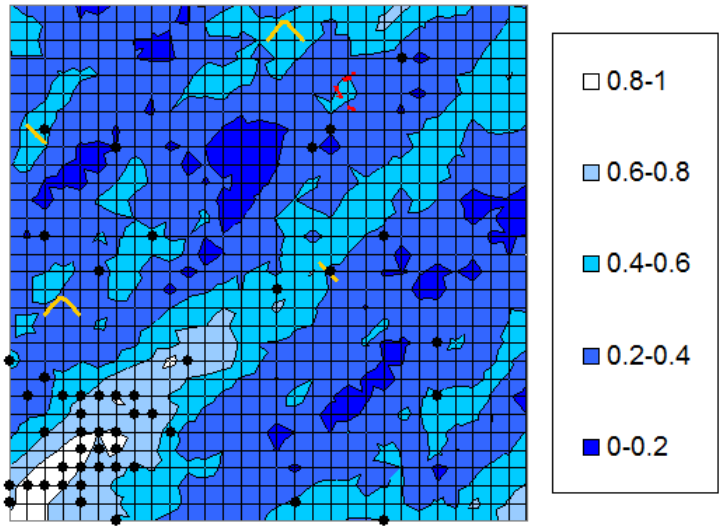


Figure 4-9: Demonstration of fuzzy-swarm algorithm operation (complex terrain)

Even through less than 7% of the possible solutions are evaluated (60 points out of a possible 900), the fuzzy swarm algorithm is often able to find the exact location of the highest point, even in complex terrains. This is highlighted in Table 11, which shows the results of ten runs of each algorithm for simple, single peak terrains, and more complex, multiple-peak terrains.

Table 11: Results of ten algorithm test runs on terrain surfaces

	Simple Terrain			***Complex Terrain***		
Algorithm Run	*Hill Climbing*	*Parallel Hill-Climbing*	*Fuzzy Swarm*	*Hill Climbing*	*Parallel Hill-Climbing*	*Fuzzy Swarm*
1	0.74	0.77	**0.95**	0.85	0.93	**0.94**
2	0.68	0.92	**0.97**	0.79	**1.00**	**1.00**
3	0.90	0.87	**0.86**	0.39	0.75	**0.79**
4	0.58	0.91	**0.97**	0.54	0.76	**0.82**
5	0.77	0.89	**0.88**	0.49	0.84	**1.00**
6	0.81	0.78	**1.00**	0.78	0.68	**0.86**
7	0.92	1.00	**1.00**	0.76	0.87	**1.00**
8	0.90	**0.96**	**0.90**	0.64	0.87	**1.00**
9	0.50	**1.00**	**1.00**	0.63	0.65	**0.94**
10	**1.00**	0.94	0.96	0.74	**0.98**	0.83

Even with simple terrains, in which hill-climbing is generally the preferred method, the fuzzy swarm algorithm performs better. This may however be results of premature-halting of the hill-climbing methods, which are only permitted to evaluate 60 points in order that these methods can be fairly compared. In more complicated terrains the difference in results is even more noticeable.

For more general optimisation problems, the nature of the search space (arrangement of contours) is normally unknown, however even if the search space is discontinuous, this methodology will generally produce a far better solution than simple hill-climbing techniques and require far fewer evaluations than traditional particle swarm, or brute-force searches. By highlighting particular areas of interest within the search space, the algorithm provides more flexibility for the user, whilst still ensuring they select a solution that demonstrates a good fulfilment of their criteria. Additionally, this algorithm is exceptionally simple to implement, making it an ideal starting point for wide variety of search/optimisation problems.

4.2.2.5 Scoring Function

An important aspect of any optimisation algorithm is determining a scoring factor. Whilst this stage can be somewhat eliminated through the use of tournament selection (Andrzej and Stanislaw, 2000) the most straightforward approach is to establish how well a particular solution fulfils a particular criteria and score it accordingly. Typically scores range from 0 to 1, although for this algorithm this is somewhat arbitrary since particle selection is assigned using a roulette-wheel selection process. Since the desired parameter in this optimisation is efficiency, it might be assumed that the score can be extracted directly; however this isn't necessarily the case. Table 12 indicates the calculated efficiencies of a thousand random possible solutions.

Table 12: Typical efficiencies of a thousand random solutions

Efficiency	*Percentage of Solutions*
<50%	2.2%
50-60%	0.8%
60-70%	3.0%
70-80%	7.3%
80-90%	47.8%
90-95%	37.0%
>95%	1.9%

From this table it is clear that if efficiency were used directly as a scoring criterion, then the majority of the latter solutions would fall closer to the 80-90% solutions, simply because there is more of them, rather than focusing on the desirable, higher scoring, 95%+ efficiency region.

For any scoring function it is generally desirable to not completely eliminate any particular solution, since the overall 'best' solution maybe close to it, although it is far less likely. Figure 4-10 gives an example of some typical scoring functions that could be used, including a simple power function, a sinusoidal function, and an exponential-power function.

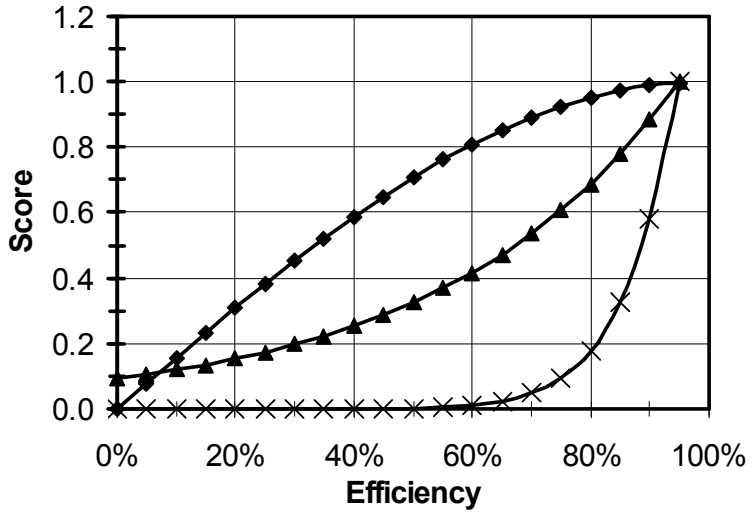


Figure 4-10: Scoring function (x = Power, ▲ = Exponential-Power, ■ = Sinusoidal)

The use of each function depends largely on the nature of the search space, and typical output parameters. Looking at Table 12, neither a sinusoidal or exponential function would sufficiently reduce the influence of lower scoring solutions such that the higher scoring solutions would be more dominant, and hence a simple power function will be used. The advantage of this function is that it is easy to modify it by adjusting the exponent of the function (scoring-power factor), which is discussed in more detail later. The effect of this scoring function on the roulette-wheel segment size in shown in Figure 4-11, which shows the efficiency of ten possible solutions together with the resulting segment size based on a scoring-power factor of 20 (each final score is raised to the power of 20).

This roulette wheel is used to determine which previous point is chosen to attract the next point. High efficiency points thus have a much higher probability of attracting further points, increasing the mapping in the local area.

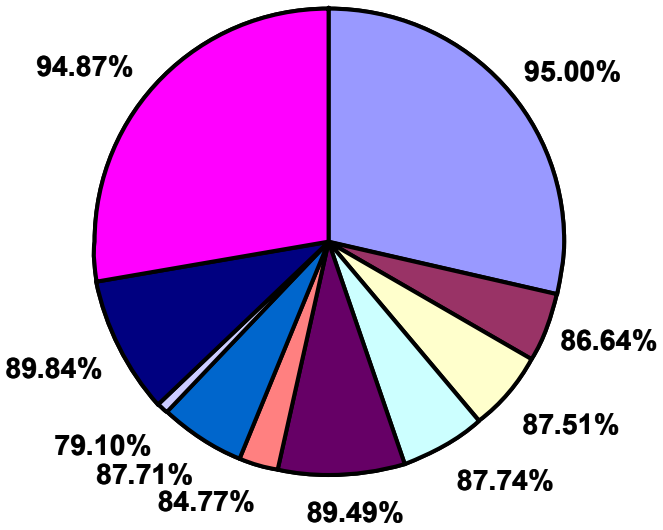


Figure 4-11: Example of roulette wheel selection for 10 solutions

This wheel will obviously change as the algorithm develops as the size of each segment is a based on a fraction of the number of solutions already evaluated.

4.2.2.6 New-Point Probability

For every step of the algorithm the dimensions generated can either be completely random, or fall close to an existing point. The specific probability of new dimensions being generated is an important parameter. Whilst in the terrain example, the probability of a point being assigned to a new location is based on a simple, discontinuous function; the probability for the actual algorithm is based on a steadily decreasing function. Experimentation with various functions indicated that the first 100 evaluations should all be new in order to 'scout' the search space, whilst after this most appropriate function is of the form:

$$P_{new} = \frac{1}{n^x}$$

Where n is the number of evaluations run, and x is a variable that can be changed to adjust the overall likelihood of new points appearing. The effect of changing x is shown in Table 13. The second column indicates approximately how many of the evaluated points were randomly generated rather than being attracted to an existing point. There isn't an absolute 'best' value for x. Larger values reduce the number of new points assigned which decreases the overall mapping of the search space, but increases the localised mapping of high-efficiency regions. Conversely, smaller values tend to cause the search space to be mapped in more detail, but reduce the chance of finding a local peak. Since the purpose of this algorithm is to determine the highest possible efficiency, a larger value of x is favourable.

Table 13: Effect of *x* on new point likelihood

x	*New points*
1	32%-35%
1.5	22-24%
2	18-20%
5	12-13%

4.2.2.7 Dimensional Limits

Assuming the operating conditions remain the same as stated previously (120Nm at 5000rpm; 62.8kW), then the dimensions of the CP-CVT need to be limited to ensure that the size and power capacity remain plausible. Approximate limits can be found by looking at Figure 2-9, which is reproduced here for ease of reference:

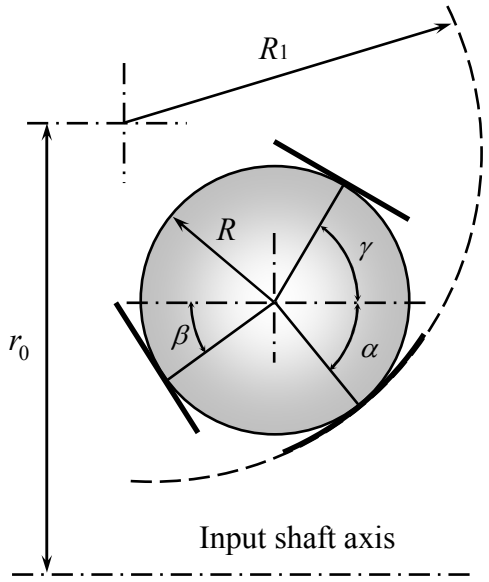


Figure 4-12: Explanation of CVT dimensions

From this figure, there are three immediately obvious physical restrictions:

1. β must be smaller than γ, else the ball element cannot be in equilibrium
2. R_1 must be larger than R, or the surface of the ball elements will be larger than the curvature of the toroidal disc
3. r_0 must be larger than R_1, of the curvature of the toroidal disc will be interfere with the input shaft.

If these limits are not included, the algorithm will attempt to evaluate mathematically impossible solutions, resulting in an error. Realistically a CP-CVT of this type and for this

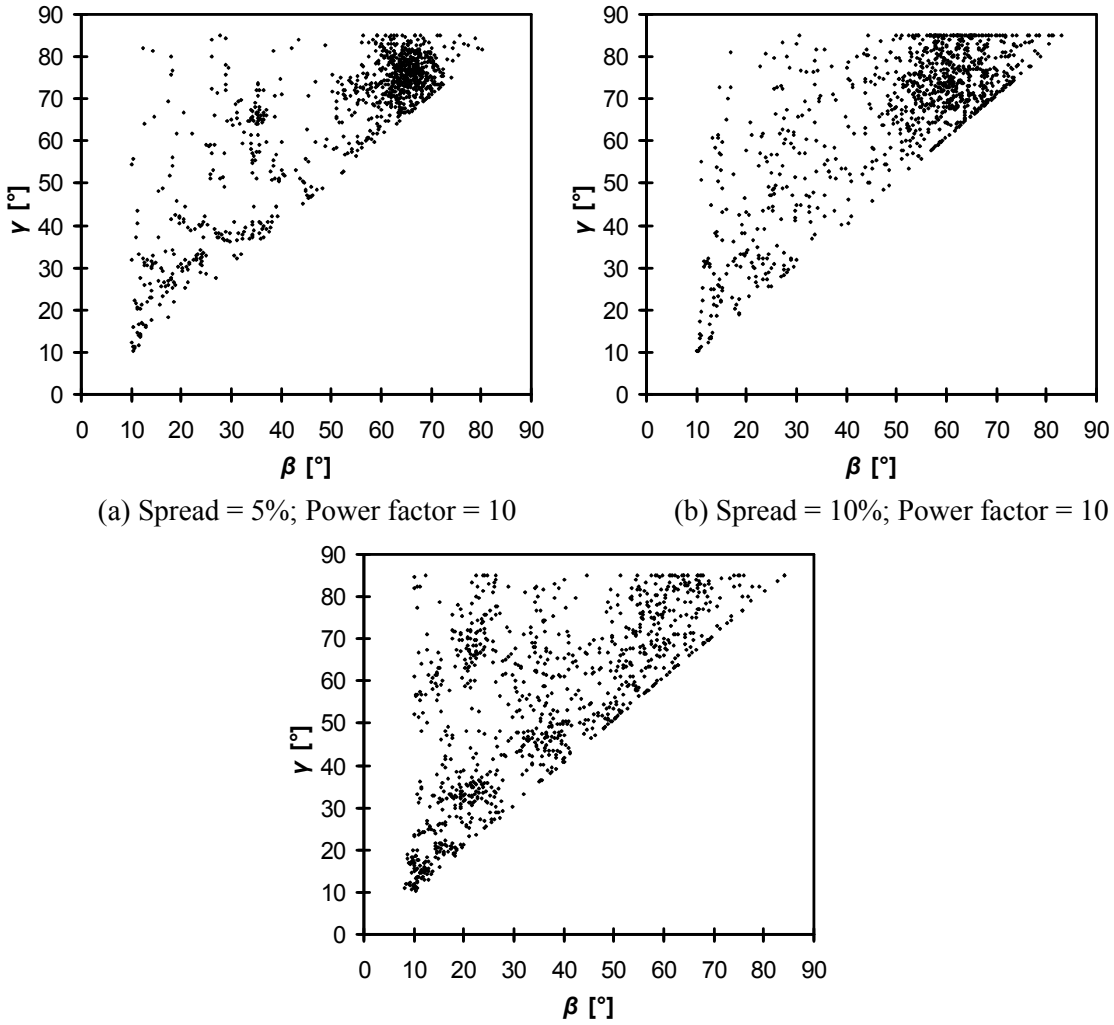
power capacity should not exceed an overall diameter of 300mm; hence r_0 should not exceed 150mm. There are several other physical limits, which will be discussed in a later section, which looks in more detail at the physical size and feasibility of the CP-CVT. Experimentation with several possibilities yielded the following domains for the input dimensions:

$$4^\circ < \beta < \gamma < 86^\circ$$
$$0.05m < r_0 < 0.15m$$
$$(40\% \times r_0) < R_1 < (90\% \times r_0)$$
$$(40\% \times R_1) < R < (90\% \times R_1)$$

These restrictions also ensure that all the dimensions remain a sensible size in relation to one another, and are only applied when producing new points. When further points are assigned, the influence of drift may cause these restrictions to be broken.

4.2.2.8 The Effect of Spread and Scoring-Power Factor

In this algorithm, the choice of scoring-power factor and spread has a large influence on the success of the algorithm in obtaining consistent and reliable results. In order to demonstrate this importance, Figure 4-13 (a-c) shows the effect of varying the spread and scoring factor on the distribution of contact angles β and γ.



(a) Spread = 5%; Power factor = 10

(b) Spread = 10%; Power factor = 10

(c) Spread = 5%; Power factor = 2

Figure 4-13: Effect of spread and power factor on algorithm convergence

Figure 4-13(a) shows the distribution for the values that were used for the algorithm, (spread = 5%, scoring-power factor = 10). This shows that the majority of the search space was at least partially mapped, whilst the area of high efficiency was found relatively quickly, and solutions generally converged in and around this area. Conversely, Figure 4-13(b) shows the effect of increasing the spread. A larger area of interest has been found with the disadvantage that a larger number of lower efficiency solutions had to be evaluated, wasting computational time. Similarly, Figure 4-13(c) shows how the importance of using a relatively high scoring-power

factor to distinguish between solutions with high and very high efficiencies. By losing this distinction, the algorithm can no longer determine which area of the search space is of particular interest, and hence dimensions are evaluated almost randomly, with no particular indication of where the best solution may lay.

4.2.2.9 Summary of Algorithm

The previous few sections have introduced a number of elements of the algorithm. In order to demonstrate how these are incorporate, Figure 4-14 shows a flow chart of the algorithm process.

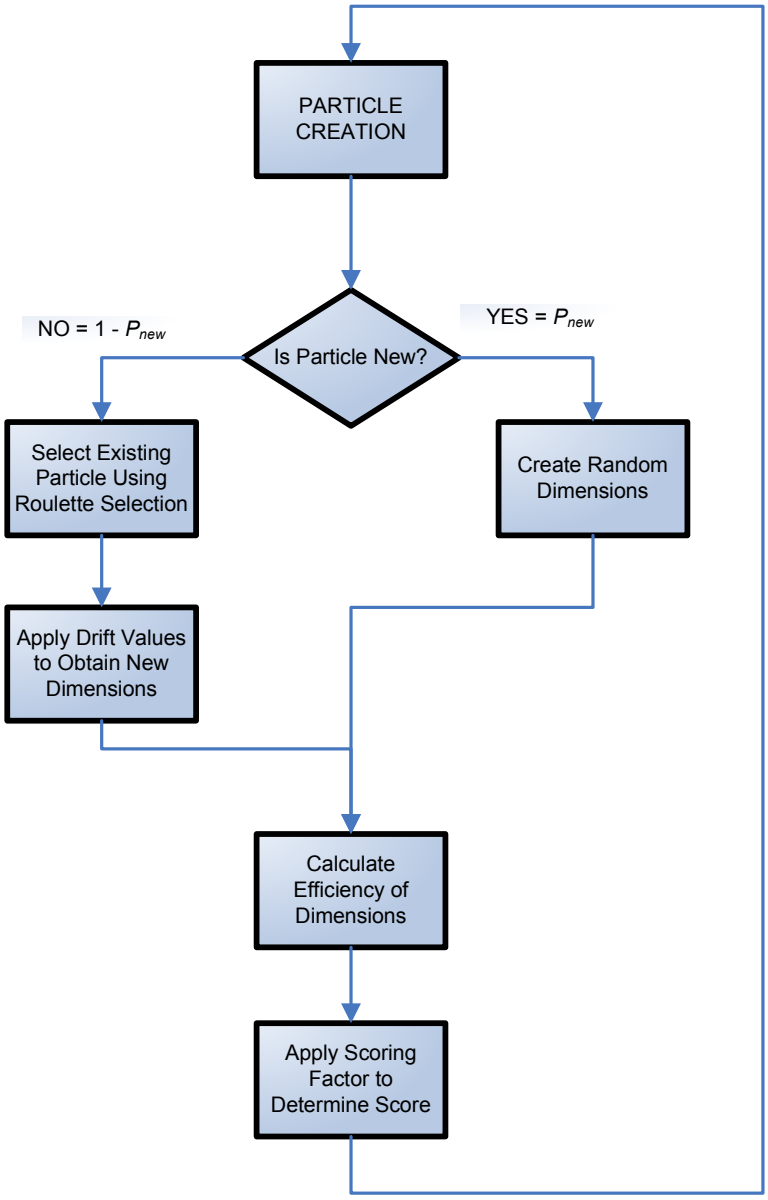


Figure 4-14: Flow chart showing process of fuzzy-swarm algorithm

4.3 Results and Discussion

4.3.1 Algorithm Results

In order to determine the robustness and consistency of the algorithm, it was repeated ten times, with each run using only a thousand evaluations of the efficiency calculations. The results of this are shown in Table 14.

Table 14: Solutions produced using optimisation algorithm

	β [°]	γ [°]	R [m]	R_1 [m]	r_0 [m]	*Overall Efficiency*
Solution 1	65.7	82.6	0.049	0.054	0.060	96.10%
Solution 2	62.5	85.6	0.043	0.048	0.054	96.22%
Solution 3	72.2	84.7	0.046	0.051	0.057	96.20%
Solution 4	70.6	81.7	0.046	0.054	0.060	96.03%
Solution 5	**67.3**	**84.5**	**0.041**	**0.045**	**0.050**	**96.26%**
Solution 6	73.3	86.0	0.036	0.048	0.054	96.08%
Solution 7	78.2	84.0	0.035	0.040	0.061	95.76%
Solution 8	64.5	82.3	0.044	0.049	0.054	96.14%
Solution 9	73.3	83.8	0.041	0.046	0.051	96.14%
Solution 10	61.8	84.7	0.052	0.058	0.065	96.07%
Average	**68.9**	**84.0**	**0.043**	**0.049**	**0.057**	**96.10%**
Variation	**11.9%**	**2.6%**	**20.1%**	**18.3%**	**12.6%**	**0.3%**

The highest overall efficiency obtained (96.26%) remains marginally lower than manual transmissions, but higher than automatic transmission and other CVT arrangements (Kluger and Long, 1999). These values could perhaps be viewed as optimistic however, as invariably actual losses are higher than those predicted by theory. In either case overall fuel efficiency would be noticeably improved through efficient engine operation.

The relatively low variations in the suggested dimensions indicate that the algorithm is generally consistent. This consistency could potentially be improved by increasing the number of evaluations, since the longer the algorithm runs, the more it learns about the nature of the search space, and hence the more the solutions would tend to converge. However given the comparatively low variation in overall efficiency, it is possible that within a particular region of the search space, there is little to distinguish between various combinations of dimensions. This finding could not be achieved through standard, single output algorithms.

The relatively low variations in contact angles perhaps indicate that they have a larger effect on the total efficiency, with only a small range of values yielding a good overall efficiency. This can be further determined through sensitivity analysis.

4.3.2 Contribution of Each Loss Source

The dimensions of the key components have a greater effect on certain sources of loss, such as the contact losses, and a lesser effect on others, such as bearing or churning losses. Table 15 shows the contribution of each loss source for all the solutions (Column 1) and for the best solution currently found (Column 2).

Table 15: Contribution of each loss source to overall efficiency

Source of Loss	*Average Contribution*	*Contribution for Best Solution*
Bearings	13.7%	18.7%
Churning	24.5%	17.0%
Creep	4.6%	13.8%
Spin	15.8%	8.4%
Cage	41.3%	42.1%

From this table it is clear that cage losses are highest. This implies that the generally the component dimensions will tend towards those that have the greatest influence on cage losses, which from Equation 3.16, implies larger values of r_0 and smaller values of R. However further investigation proved this is not actually the case, and generally cage losses are not particularly dependent on component dimensions, hence cage losses simply have to be accepted, and to a certain extent, ignored in terms of dimensional optimisation. Cage losses can be significantly reduced through the use low friction coatings, specially designed cages that reduce the Hertzian contact pressure, or the use of rollers that ensure the contact has an element of rolling instead of being pure sliding. Whether or not this is justifiable in terms of cost and maintenance would depend on the particular application.

Churning losses are also fairly high, which maybe a result of inaccurate theoretical calculations, for which there still remains a lot of conflict in existing literature, however without further experimentation it is difficult to confirm this. These losses could potentially be eliminated in a similar way to existing dry-sump engines rather than using dipped lubrication. However as with engines, this can significantly increase costs, and add an additional loss in the form of the

energy required to re-circulate the traction fluid, making this perhaps an undesirable modification.

Interestingly for average dimensions spin losses are much higher than creep losses, whilst for better solutions the opposite is true. Potentially this could be because the best solutions found tend to reduce the normal forces, reducing cage losses, which have the highest influence on overall efficiency. However, by reducing the normal forces, the traction coefficient between the contacts is increased, increasing slip, and hence creep losses.

For this particular CVT, bearing losses are not particularly high. Despite this, bearing losses could be reduced through some adaptation of the nature of the discs to balance the axial shaft loads, hence reducing the load applied to the bearings.

4.3.3 Sensitivity of Input Dimensions

In order to assess the local sensitivity of the dimensions, each individual dimension was varied between the minimum and maximum values found in Table 14, whilst the non-varied dimensions were set to the average values shown. Figure 4-15 shows the effect of β and γ on overall efficiency and total contact losses (spin and creep).

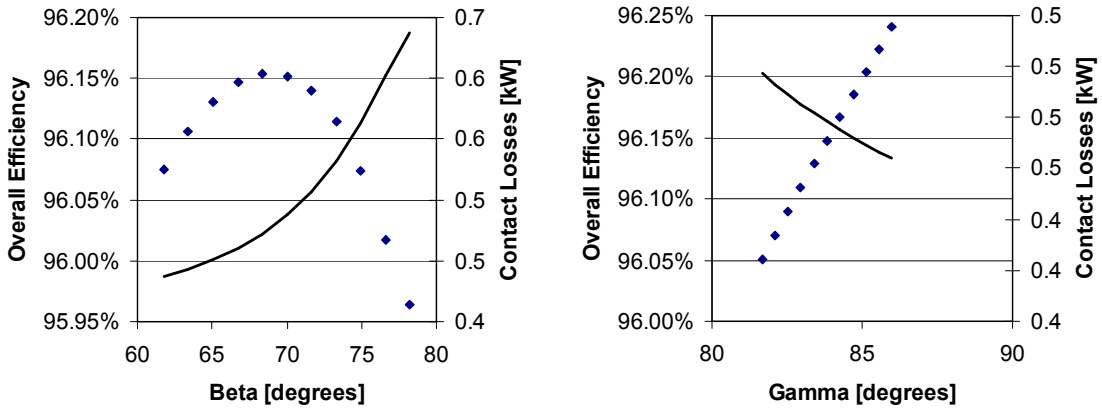


Figure 4-15: Influence of β and γ on overall efficiency (symbols) and contact losses (lines)

Looking at the effect of β on overall efficiency, it is clear there is a very defined and prominent local maximum point, which occurs at 69°. Interestingly, this is not the value of β that provides the lowest contact losses. Conversely, there is no local peak in efficiency for the value of γ, instead the results indicate that it should take as large a value as possible, hence the reason that dimensions offered for γ in Table 14 tend towards the imposed limit of 86°. This indicates that

conical output disc's surface should be as parallel to the input shaft as possible, however this would significantly increase the length of the disc and hence the overall length of the CVT.

Similarly, Figure 4-16 indicates the effect other dimensions on the overall efficiency and contact losses.

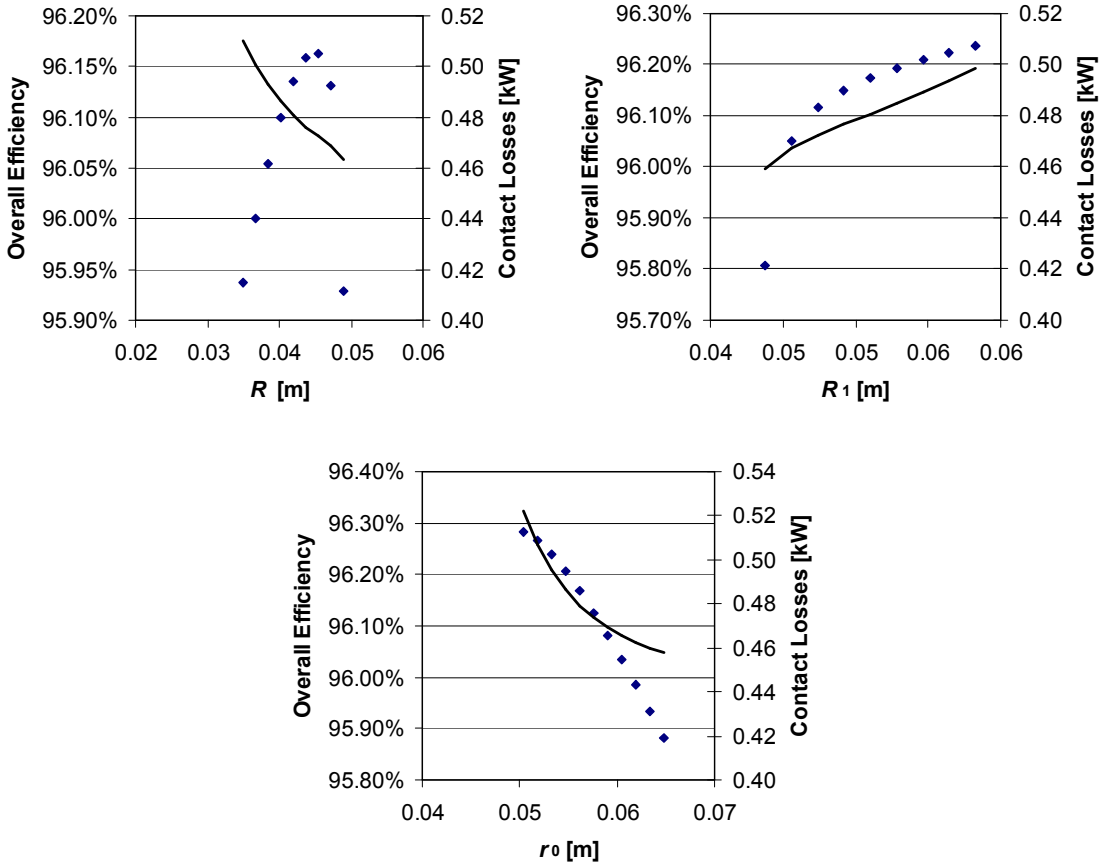


Figure 4-16: Influence of R, R_1 and r_0 on efficiency (symbols) and contact losses (lines)

As was the case with β, the radius of the ball elements R, has a well defined 'best' value. Larger diameter ball elements would significantly decrease the Hertzian pressure at the contact, decreasing contact losses, (as shown by the solid line in the graph for R). However this comes at a cost of an increased value of r_{min} and hence the immersion depth, increasing churning losses, hence the reason the overall efficiency decreases as the value of R increases beyond a limit.

The values of R_1 and r_0 are closely related. It was previously established that R_1 must be smaller than r_0, and yet the highest efficiency can be achieved by having larger values of R_1 and smaller values of r_0, despite the increased contact losses this causes.

By setting each of the dimensions to their ideal values illustrated in Figure 4-15 and Figure 4-16, the maximum overall efficiency that could be achieved was 96.40%. This was achieved using the dimensions shown in Table 16.

Table 16: CVT dimensions demonstrating the highest overall efficiency

Parameter	*Value*
β	68.4°
γ	86°
R	0.044m
R_1	0.049m
r_0	0.050m

4.3.4 Discussion of Best Solution

4.3.4.1 Influence of α

It was stated previously that a single, representative value of α was used for the efficiency calculations (17°) in order to reduce the number of calculations required. The actual influence of α on transmission ratio and overall efficiency is shown in Figure 4-17.

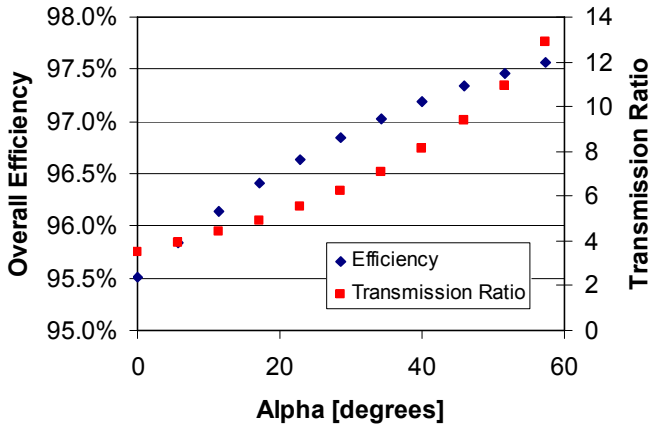


Figure 4-17: Influence of change in alpha on transmission ratio and overall efficiency

This figure shows that efficiency of the CVT changes significantly with α, and hence the transmission ratio. It might be reasonable to assume that a more representative value of α would be 30°, (approximately half of the maximum value of 1 radian), however Kluger and Long (1999) found that the vehicles typically spend more in either first gear (moving from rest), or maximum gears (whilst cruising).

4.3.4.2 Effect of Input Speed and Torque

The stated efficiency could theoretically be raised through increasing the running torque of the CVT, whilst decreasing the running speed, hence keeping the power transmitted the same. This could be achieved through additional gearing placed between the power source (engine) and the transmission, whilst the final drive ratio (differential) which is placed after the transmission in the drive train, could be adjusted to compensate for this additional ratio. Although this would add another source of loss (gearing losses), this might be offset by an increase in CVT efficiency as shown in Figure 4-18.

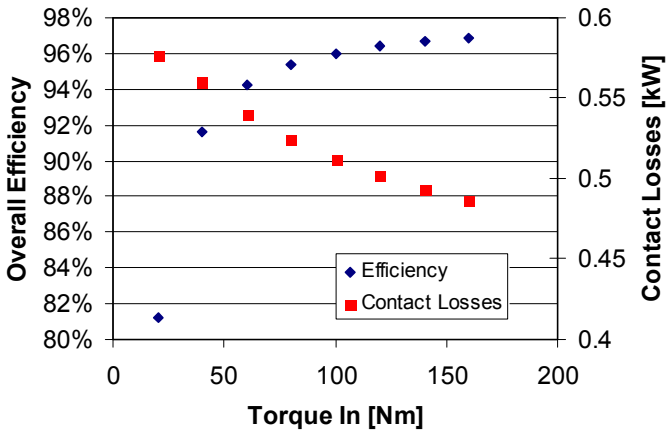


Figure 4-18: Influence of pre/post-gearing on contact losses and overall efficiency

From this figure it is clear that by increasing the running torque, efficiency is improved and contact losses are reduced. However whilst efficiency may increase this comes at the cost of a significant increase in Hertzian pressure, which has a large influence on wear and overall life expectancy of the CVT. Instead of increasing the running torque, a more attractive solution maybe to reduce the overall size of the CP-CVT, provided the Hertzian pressure remained below the theoretical limit of 2GPa (Lee et al., 2004). Reducing the size would reduce manufacturing costs, mass, and some of the losses such as churning, making it a far better alternative to pre/post gearing.

4.3.5 Feasibility of Solution

4.3.5.1 Ball Element

In order to quickly assess the feasibility of any given set of dimensions, a simple program was written that plots the layout of each of the key components. To do this, each component's extremes (the input and output discs, and ball element) are expressed as a series of coordinates. Since all components are in contact with the ball element, and other previously calculated variables are based around the input shaft, the easiest location to place the origin of the Cartesian frame of reference is with the *x*-axis running through the centre shaft and the *y*-axis through the centre of the ball element in its initial position, as shown in Figure 4-19.

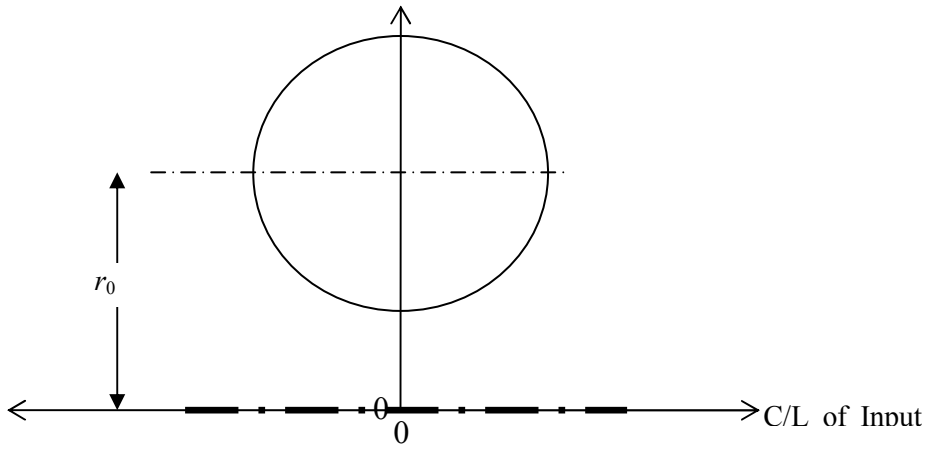


Figure 4-19: Position of Origin for Plot

Based on this system, the centre of the circle has co-ordinates (0, r_0), whilst the equation that dictates the outline of the circle is:

$$(R\cos\theta, r_0 + R\sin\theta)$$
$$for: 0 \le \theta \le 2\pi$$

4.3.5.2 Conical Input Disc

The conical input disc can be plotted as shown in Figure 4-20 using the following coordinate points:

Point	***x***	***y***
(1)	$-R\cos\beta - L_i$	0
(2)	$-R\cos\beta - L_i$	$r_0 - R\sin\beta$
(3)	$-R\cos\beta$	$r_0 - R\sin\beta$
(4)	$-R\cos\beta + (R_1\text{-}R)\sin57\tan\beta$	$r_0 - R\sin\beta - (R_1 - R)\sin57$
(5)	$-R\cos\beta + (R_1\text{-}R)\sin57\tan\beta$	0

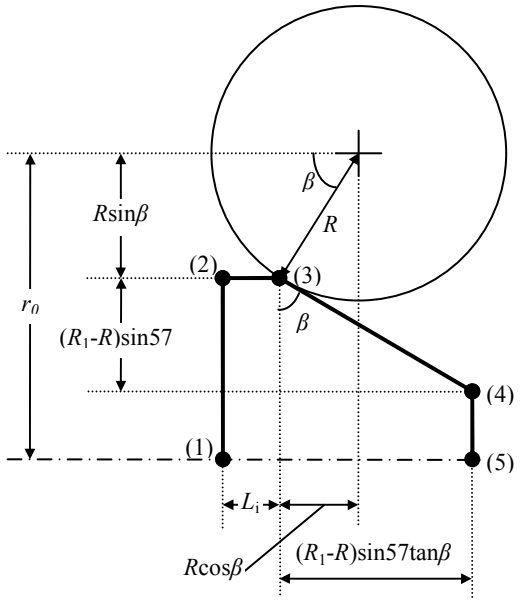


Figure 4-20: Coordinates of Input Conical Disc

4.3.5.3 Coordinates of Conical Output Disc

The conical output disc can be plotted as shown in Figure 4-21, using the following coordinate points:

Point	*x*	*y*
(1)	$R\cos\gamma + (R_1\text{-}R)\sin57\tan\gamma + L_{oh}$	$r_0 + R\sin\gamma + L_{ov}$
(2)	$R\cos\gamma$	$r_0 + R\sin\gamma + L_{ov}$
(3)	$R\cos\gamma$	$r_0 + R\sin\gamma$
(4)	$R\cos\gamma + (R_1\text{-}R)\sin57\tan\gamma$	$r_0 + R\sin\gamma - (R_1\text{-}R)\sin57$
(5)	$R\cos\gamma + (R_1\text{-}R)\sin57\tan\gamma + L_{oh}$	$r_0 + R\sin\gamma - (R_1\text{-}R)\sin57$

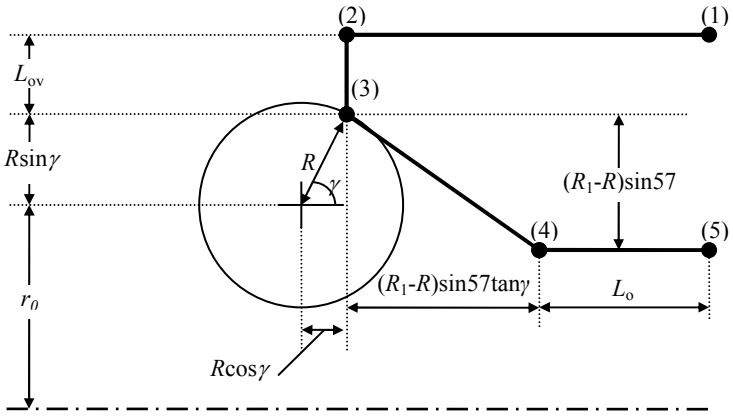


Figure 4-21: Coordinates of Output Conical Disc

4.3.5.4 Coordinates of Toroidal Input Disc

The coordinates for the input toroidal disc are more complex since it consists of a mixture of straight and curved lines. The curve can be expressed in a similar way to the ball element; as an arc with centre at (R–R_1, r_0). The coordinates of the arc would hence be:

$$\left([R - R_1 + R_1 \cos\alpha],[r_0 - R_1 \sin\alpha]\right)$$
$$for: 0 \geq \alpha \geq 57$$

Whilst straight lines would simply connect this to the centre line of the shaft. By calculating the movement of each component, a similar plot can be made for when $\alpha = 57°$.

4.3.5.5 Plot of Dimensions derived from Algorithm

The resulting plot of the dimensions stated in Table 16, which shows both extremes ($\alpha = 0°$ and $\alpha = 57°$) is shown in Figure 4-22.

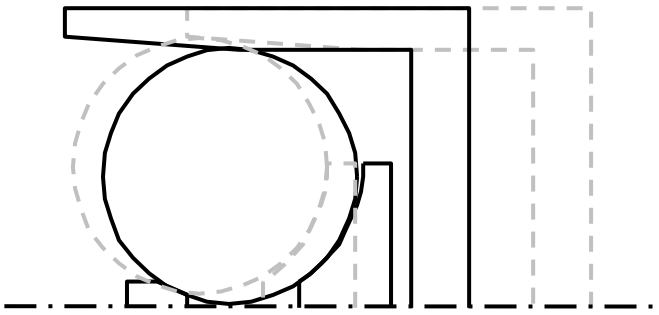


Figure 4-22: Graphical plot representation of layout of CP-CVT

From this figure it clear that these set of dimensions would not be feasible. Firstly, the value of r_0 is too small, and hence there is interference between the ball elements and the input shaft. Furthermore, the length and movement required of the output disc is too large, meaning the overall casing length of the CVT would be far too big, even though the majority of the internal volume is empty space, which is not a particularly efficient design.

4.4 Conclusions

This chapter has presented a useful method of efficiency optimisation for the CP-CVT. Whilst perhaps not as accurate as other methods, the comparatively simple efficiency calculations, including unique approaches to more complex EHD contact behaviour, allowed a very practical comparison of various dimensional combinations of traction drive components. Although the efficiency values stated could be considered representative, rather than absolute, they do show the potential improvement that can be made in transmission efficiency compared to the component dimensions offered in previous CP-CVT literature.

It was found that a large proportion of the losses of the CP-CVT arise from the ball separator (cage), however since this specific part has not been fully designed yet, it is difficult to ascertain an absolute value for the effective traction coefficient and hence losses. As stated previously, the value used (0.02), is optimistic, however by increasing this, cage losses completely dominate the overall efficiency, and hence fitness score. The problem with this is that other sources of loss are somewhat ignored during optimisation and the resulting dimensions simply favour those that yield the lowest cage losses.

Although representative, the overall calculated efficiency of the CP-CVT is a significant improvement on current automatic transmissions and a slight improvement over alternative CVT designs. One of the reasons for this is the absence of side-slip within the contact and the lack of an actuator to control the transmission ratio and provide the normal force, both of which decrease overall efficiency. This is highlighted by the results of Kluger and Long (1999), whom found similar contributions to loss for both creep/spin (16%), and bearings (10%) for a different traction drive design, but showed that the use of a hydraulic pump to generate the normal forces contributed to 67% of the total losses.

The stated efficiency could theoretically be raised through pre and post-gearing, increasing the torque applied to the CP-CVT, whilst decreasing the running speed, which has a large influence on a number of the losses. However, whilst efficiency may increase, this comes at the cost of increased contact forces in order to maintain a low traction coefficient, which increases Hertzian pressure and reduces the overall life expectancy of the device, making it an unattractive solution.

The algorithm itself performed very well, despite its relative simplicity and ease of implementation. Further detailed analysis of the high-efficiency local search space indicated that the maximum improvement that could be achieved in efficiency was only 0.14%. This indicates that although a wide range of dimensions were offered, they are all reasonable solutions, any of which could be considered acceptable, and in some cases favourable due to

other considerations such as mass or size. Further refinements could be made to improve the consistency of results, depending on the type of problem it is applied to. Currently each run of the algorithm uses only a thousand evaluations of the efficiency calculations, which is significantly less than is required for genetic algorithms or hill-climbing techniques. This makes it a very attractive 'first-step' in optimisation before more advanced techniques are later applied with multi-objective criteria.

One of the problems with this type of optimisation is the specific nature of the torque and power requirements. This means that certain solutions may have high losses, but only because they are much larger and hence have the capability of withstanding a significantly higher power. When the losses are taken as a percentage, it might be found that certain solutions that have been disregarded are actually favourable. Furthermore, without considering other properties such as overall size and mass, certain solutions that seem favourable may later be found to be unrealistic. In order to overcome this, all properties of the CP-CVT need to be optimised simultaneously.

CHAPTER 5: MULTI-CRITERIA DIMENSIONAL OPTIMISATION

5.1 Introduction

5.1.1 Chapter Summary

Having determined that the efficiency of the CP-CVT can be improved through geometric optimisation, it is necessary to simultaneously analyse and optimise all the technical requirements together. One of the greatest problems with the dimensional optimisation of the CP-CVT is the inter-dependencies of all of these requirements. As in the previous chapter, there are five input dimensions, all of which have a strong influence on the overall properties of the CP-CVT. There is not a single optimum value for each dimension, instead each one is limited and affected by the others. Previous research by Cretu and Glovnea (2006) found that there is no single 'best' combination of dimensions that are universally applicable. Conversely each new application requires a different set of dimensions to fulfil specific criteria. What makes this problem difficult is the vast number of combinations that are possible. Assuming a resolution of 1mm and 0.5°, there are approximately $8x10^{11}$ possible combinations, hence choosing the optimum solution for any particular problem can be very time consuming.

There are several possible techniques that could be used to achieve this multi-criteria optimisation. One possible method would be to start with a particular set of dimensions then improve each one in isolation as necessary until a better solution is found. The problem with this approach is that the resulting solution is influenced by the initial dimensions selected. An alternative solution is to systematically test each possible combination (brute force), however this requires vast computational time, and would have to be repeated each time there is a change in the operating conditions, such as engine power, or changes to the perceived importance of each criteria as derived from the House of Quality. A better method of optimisation is the use of multi-criteria optimisation algorithms, such as the one discussed in the previous chapter, or a genetic algorithm.

5.1.2 Technical Requirements

Based on the design theories discussed in Chapter 3, the technical requirements have already been determined:

- Device efficiency
- Transmission ratio range
- Torque capacity
- Mass
- Length
- Diameter
- Output power variation

Each of these specific and measurable requirements will form the basis for the evaluation functions that will assess the quality of each particular set of dimensions in fulfilling the customer demands, based on the weighting factors derived previously from the House of Quality.

5.1.2.1 Device efficiency (η)

Detailed information regarding the calculation of efficiency has already been presented in Chapter 4. These calculations are repeated here, with the only change being an increase in the predicted traction coefficient between the ball separator (cage) and the ball elements. Previously an optimistic value of 0.02 was used to ensure that cage losses did not completely dominate the overall losses. Since the individual contributions are no longer of interest, a more realistic value of 0.06 will be used instead.

5.1.2.2 Normalised Transmission Ratio Range (i)

For any transmission system one of the most important properties is the range of ratios the transmission is able to offer. Toroidal CVTs are generally able to provide a better range or ratios than standard, step-gear transmissions. In order to provide a single comparable value, a normalised transmission ratio is used for evaluation. This is simply the maximum transmission ratio divided by the minimum transmission ratio. The equation for calculating the transmission ratio of the CP-CVT as a function of α has already been derived (Equation 2.14); hence the normalised transmission ratio is simply:

$$i_{norm} = \frac{i_{\max}(\alpha = \max)}{i_{\min}(\alpha = \min)} \tag{5.1}$$

5.1.2.3 Torque Capacity (T)

Although it is perhaps more common to measure a transmission device's performance in terms of power capacity, the limiting factor for toroidal-type CVTs is normally the Hertzian contact pressure and the lubricant's traction coefficient. Both of these factors limit the torque capacity, and hence this parameter is a better indicator of the device's performance.

A significant amount of research has shown the importance of reducing Hertzian pressure in order to improve the durability of toroidal CVTs. Although there is not a single universal value for the maximum Hertzian pressure, Lee et al. (2004) found that fatigue life is considerably increased for when the pressure is below 2GPa.

Both the output conical disc and toroidal disc have surface curvatures that are concave in at least one axis, significantly reducing the contact pressure, whilst the conical input disc has only convex curvature, hence the peak Hertzian pressure invariably occurs at the conical input disc (this has been confirmed with numerous test calculations, not shown). By rearranging the classic Hertzian pressure equation for elliptical contacts (discussed previously), it can be shown that the normal force at the contact between the conical input disc and the ball element (N_B) is a function of the peak Hertzian pressure (P_0), as shown in Equation 5.2.

$$N_B = \frac{P_0^{\,3}\pi^3 R_\sigma^{\,2}}{6\overline{E}^2} \tag{5.2}$$

Rearranging the normal force equations derived in the previous example calculations yields the follow expression for the relationship between N_B and N_C:

$$N_B = N_C \frac{(\sin\gamma + \cos\gamma\tan\alpha)}{(\cos\beta\tan\alpha + \sin\beta)} \tag{5.3}$$

Furthermore, a relationship can be derived for the relationship between the maximum torque output (T_{out}) and N_C based on a maximum traction coefficient (μ), which has been found to be 0.045 (Fuchs and Hasuda, 2004).

$$N_C = \frac{T_{out}}{\mu\ [r_0 - (R_1 - R)\sin\alpha + R\sin\gamma]} \tag{5.4}$$

Hence by combining Equations 5.2-5.4, a relationship can now be derived for the relationship between the maximum allowable torque for a given Hertzian pressure limit and maximum traction coefficient:

$$T_{out} = \frac{P_0^{\,3}\pi^3 R_\sigma^{\,2}}{6\overline{E}^2}\left[\frac{(\cos\beta\tan\alpha + \sin\beta)}{(\sin\gamma + \cos\gamma\tan\alpha)}\right] \times \mu\ [r_0 - (R_1 - R)\sin\alpha + R\sin\gamma]$$

It is generally more useful to indicate the maximum torque the CP-CVT is able to handle at the input, which can be found relatively easily using the transmission ratio as shown in Equation 5.5

$$T_{in} = \frac{T_{out}}{i} \tag{5.5}$$

5.1.2.4 Indicative Mass (m)

Mass is an especially important property in automotive applications. A large mass, especially in the front of vehicle can negatively affect fuel efficiency, performance and handling. The majority of the mass within the CP-CVT will be comprised of the primary components (input discs, output discs, and ball elements). The calculations conducted exclude factors such as casing, traction fluid, loading system, etc, and hence is a relative, not an absolute measurement, however it does provide a useful comparison between various solutions.

The derivations of the calculation of the volume of each of the key components are relatively complex and are presented in Appendix A.3. A summary of the results of these derivations is shown here:

The conical input disc volume:

$$V_{ci} = \tfrac{1}{3}\pi \tan\beta\left([r_0 - R\sin\beta]^3 - [r_0 - (R_1 - R)\sin\alpha - R\sin\beta]^3\right) \tag{5.6}$$

The conical output disc volume:

$$V_{co} = \pi\tan\gamma\left((r_0 + R\sin\gamma + 0.01)^2(R_1 - R)\sin\alpha - \tfrac{1}{3}(r_0 + R\sin\gamma)^3 + \tfrac{1}{3}[r_0 - (R_1 - R)\sin\alpha + R\sin\gamma]^3\right) \tag{5.7}$$

Toroidal disc volume

$$V_{tr} = \pi[\tfrac{1}{12}R_1^3(8 - 9\cos\alpha + \cos 3\alpha) + r_0 R_1^2(\tfrac{1}{2}\sin 2\alpha - \alpha) + r_0^2(R_1 - R_1\cos\alpha + 0.01)] \tag{5.8}$$

Where α is measured in radians.

Total volume of ball elements:

$$V_{be} = n\tfrac{4}{3}\pi R^3 \tag{5.9}$$

Where n is the number of ball elements (typically assumed to be four).

The total mass can now be calculated using the density of the disc and ball material (ρ):

$$m = \rho\left(V_{ci} + V_{co} + V_{tr} + V_{be}\right) \tag{5.10}$$

5.1.2.5 Indicative length (l)

The shape of the CP-CVT is roughly cylindrical, and hence the overall size can be characterised by the significant length and diameter. Calculating a reasonable indicative value for the length is not easy since several of the parts could be classified as the extremes depending on their particular geometry, as shown in Figure 5-1.

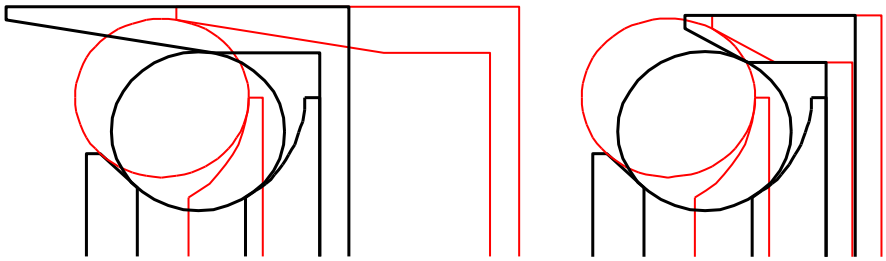


Figure 5-1: Example of different length extremes

In the left figure, the extreme point towards the input end (left-hand side) of the CVT is the conical output disc, whilst in the right figure, it is the ball element. In both cases, the extreme point towards the output shaft (right-hand side) is the connection arm of the conical output disc. Taking a coordinate system with x=0 at the centre of the ball element when α=0, the following can be stated:

$$l = x_2 - x_1 \tag{5.11}$$

Where:

$$x_1 = \min\left(-R, ((R_1 - R)\sin 57)(\tan\beta - \tan\gamma) + R\cos\gamma\right)$$
$$x_2 = R\cos\gamma + (R_1 - R)\sin 57\tan\gamma + R(\cos 57 - \cos\gamma) + R_1(1 - \cos 57)$$

Typically the total length of the CP-CVT, including casing and bearings, etc, will be approximately 70-100mm greater than this value of indicative length.

5.1.2.6 Indicative diameter (d)

The diameter of the CP-CVT has been determined to be a function of the furthest part from the centre of the shaft, which is typically the output disc. An indication of the diameter of the CVT can hence be found from Equation 5.12.

$$d = r_0 + R \tag{5.12}$$

The total diameter of the CP-CVT including casing will be obviously be at least double these radii, and hence this is only representative, rather than absolute.

5.1.2.7 Power Variation (p.v)

The variation in power is an important aspect of the principle of the CP-CVT, especially for automotive purposes, since it enables the engine to operate within a fixed regime for longer periods of time, increasing engine efficiency. For an engine to easily remain in the same regime, the load applied to it must be constant. In order to achieve this, the torque input to the CVT (T_{in}) must be as constant as possible. If it is assumed for the moment that the torque input is constant across all values of α, then from this a torque output can be calculated simply by using the transmission ratio. For a fixed ball-screw coupling, this torque output has a corresponding value of axial force (F_C) that is produced by the ball-screw, and therefore a corresponding value of F_A that is required to maintain the position of the ball elements in equilibrium (as shown in Equation 2.20). If it is assumed that variation in power is analogous to variation in torque input, then it can be stated that power variation is controlled by the value of F_A. By extension, if a simple linear spring is used ($F= k.dx$), then the variation in power can thus be indicated by the relationship between the horizontal displacement of the toroidal disc (dx) and the change in force required to maintain a constant torque input (dF_A).

If a simple spring is used then ideally there will be a linear relationship between the displacement of the toroidal disc and the force required. It is a variation from this linear relationship that causes a fluctuation in the power output, thus requiring a change in the torque output from the engine. An indication of the power variation can thus be found by using linear regression to determine the correlation coefficient (r) of the relationship between dx and dF_A:

$$r = \frac{n\sum\left(dx \times dF_A\right) - \sum dx \sum dF_A}{\sqrt{\left(n\sum\left(dx^2\right) - \left(\sum dx\right)^2\right) \times \left(n\sum\left(dF_A{}^2\right) - \left(\sum dF_A\right)^2\right)}} \tag{5.13}$$

Where $dx = dx(\alpha) = (R_1 - R)\left(1 + \sin\alpha\tan\beta - \cos\alpha\right)$

And $dF_A = dF_A(\alpha)$

For $0 < \alpha < 57^\circ$

One of the advantages of using regression is that the scale of either value is irrelevant. This means that as long as another property is proportional to dF_A then it can be substituted. This also implies that the actual operational torque of the device has no affect on its power variation. Hence:

$$F_A = f(F_C)$$

Or more specifically:

$$F_A = N_A \cos\alpha = F_C \frac{\cos\alpha}{\cos\gamma} \frac{(\sin\gamma - \cos\gamma\tan\beta)}{(\cos\alpha\tan\beta + \sin\alpha)} \tag{5.14}$$

But, for fixed ball screw properties:

$$F_C \propto T_{out}$$

Additionally it follows that for a fixed T_{in}:

$$F_C \propto i$$

Hence:

$$F_A \propto i \cos\alpha \frac{(\sin\gamma - \cos\gamma \tan\beta)}{(\cos\alpha \tan\beta + \sin\alpha)} \tag{5.15}$$

Whilst for the purposes of regression these can be considered identical. Hence the power variation value (*p.v*) is simply the coefficient of correlation between the force required at the toroidal disc and its relative movement. This is perhaps explained better through the worked example shown in Appendix A.4.

5.1.3 Example Calculation

In order to better demonstrate how each of the stated parameters are calculated, the original dimensions shown in Table 9 are used (β = 45°; γ = 75°; R = 0.06m; R_1 = 0.09m; r_0 = 0.12m) to calculate each of the stated parameters. The full calculations are shown in Appendix A.4, whilst a summary of the results is shown in Table 17.

Table 17: Summary of calculated parameter for dimensions shown in Table 9

Parameter	*Calculated Value*
Normalised transmission ratio	2.630
Maximum input torque	181Nm
Indicative mass	56.7kg
Indicative length	0.228m
Indicative diameter	0.18m
Overall efficiency	93.3%
Power variation coefficient	0.382

The results shown in Table 17 highlight the fact that there has been little attempt to optimise the CP-CVT thus far. Several of the calculated parameters would be unacceptable for a typical automotive transmission. Whilst the normalised transmission ratio, and maximum input torque are acceptable, the indicative mass (which consists only of the key components) is too high. Furthermore, the overall efficiency shown (dominated by increased cage losses) should be improved in order to compete with other traction drive designs.

5.2 Methodology: Solution Evaluation

5.2.1 Evaluating Solutions

Both genetic algorithms and the fuzzy-swarm process developed previously require that each set of dimensions is scored according to its ability to fulfil the desired criteria. One method of achieving this is the use of a Pareto efficiency-type function. For any multi-criteria problem, there exist situations where changing one of the input parameters may improve the solution with respect to a single criterion without worsening another. A Pareto optimal point is the point at which this is no longer possible. Formally the problem of dimensional optimisation of the CP-CVT can be written in quasi-Pareto terminology thusly:

If the dimensions β, γ, R, R_1, r_0 are considered as a vector variable $\mathbf{x}$*, which form an optimisation vector function* $f(\mathbf{x})$*, then the vector* $\mathbf{x}_u$ *can be considered to be Pareto-optimal only if there is no such vector* $\mathbf{x}_v$ *for which* $f(\mathbf{x}_v) > f(\mathbf{x}_u)$*, assuming the purpose is to maximise* $f(\mathbf{x})$*.*

In reality the vector function $f(\mathbf{x})$ is a compound function:

$$f(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_7(\mathbf{x}))$$

Where each of the functions $f_1(\mathbf{x})$ - $f_7(\mathbf{x})$ represent the relative fulfilment of each of the seven technical requirements discussed previously. The relative fulfilment functions are calculated by taking the scoring function of each requirement (s) and multiplying it by its relative weighting (w), hence:

$$f(\mathbf{x}) = (s_1(\mathbf{x})w_1 + s_2(\mathbf{x})w_2 + \ldots s_7(\mathbf{x})w_7)$$

In summary, in order to derive a Pareto function in which each individual function does not hold equal weighting, the follow three main factors must be determined:

1. The specific value of each criteria (mass, ratio range, efficiency, etc)
2. The fitness score of each criteria relative to a target value
3. The relative weighting of each criteria

The value of each criterion can be determined from the equations derived in the section 5.1.2; these can then used to determine the fitness score of each criterion in meeting the technical requirements.

5.2.2 Fitness Score function

It is widely considered that the hardest task in multi-criteria optimisation is assessing the fitness of each solution (Whitley, 2001). A fitness function is thus required to automatically assign a score to each individual technical requirement. This function must have the following features:

1. The resulting scores must be consistent: *This is required to ensure that each technical requirement has equal weighting (before the relative weighting is applied).*
2. The scores must vary sufficiently between solutions: *The function must ensure that the range of the scores forces better solutions to be dominant.*
3. The maximum score attainable should be at the target value: *Solutions that exceed the target value should not be favoured.*
4. Solutions that are unacceptable should have no score: *If a solution is deemed to be completely unacceptable it should not influence the overall fitness score at all*
5. Scores should vary between 0 and 1: *Although this is not strictly necessary, it does mean that when each fitness score is multiplied by its relative weighting, the overall maximum score attainable will also vary between 0 and 1, and thus provide a direct indication of how well each set of dimensions fulfils customer requirements*

Bai and Kwong (2003) followed a similar principle using the concept of constraints rather than targets, which stated that:

- s is 0 if the constraints are strongly violated (far from the target)
- s is 1 if the constraints are satisfied (the target is met)
- s should increase systematically from 0-1 between these conditions

Given that for some requirements a lower value is desired, whilst for others higher values are preferable, a single function cannot be used for every technical requirement. Until now the functions discussed have been in general terms, however to proceed further it is necessary to look specifically each function. These can be broadly divided into three categories as follows:

- Relative maximised target: Characteristics for which it is preferable to have higher relative values (Normalised transmission ratio, Maximum input torque)
- Relative minimised target: Characteristics for which it is preferable to have lower relative values (Indicative mass, Indicative length, Indicative diameter)
- Absolute maximised target: Characteristics for which the minimum and maximum values are approximately known for which it is desirable to have a larger value (Overall efficiency, Power variation coefficient)

5.2.2.1 Relative Maximised Targets

For relative maximised target requirements, such as mass, there is a target value in mind that it is desirable to attain. Any further improvement beyond this target is unnecessary, whilst a solution that yields a value significantly less than this target is unacceptable. If it is assumed that less than half of the target value is unacceptable, and hence would score 0 and that anything beyond the target will always score 1, then a typical linear function for these characteristics is shown in Figure 5-2 (solid line). In order to differentiate better between different results, this linear function can be squared to further reduce the score of those solutions that are below the target (dashed line).

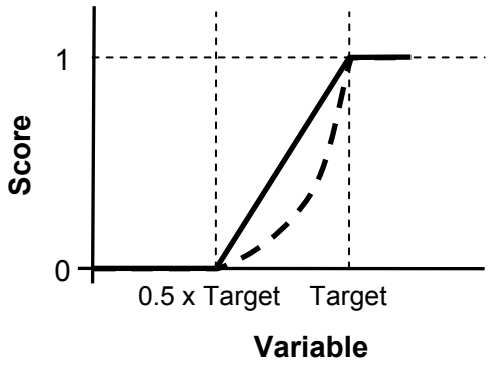


Figure 5-2: Scoring functions for relative maximised target variables

For $0.5t < v(\mathbf{x}) < t$, the linear function is:

$$s(\mathrm{x}) = 2t^{-1}v(\mathrm{x}) - 1 \tag{5.16}$$

Whilst the squared function would simply be:

$$s(\mathrm{x}) = (2t^{-1}v(\mathrm{x}) - 1)^2 \tag{5.17}$$

where t is the target value, and $v(\mathbf{x})$ is the value calculated for a particular set of dimensions.

In reality there are three distinct regions, hence for relative maximised target, the complete scoring function would be:

$$s(\mathrm{x}) = \begin{cases} 1 & \text{if} \quad v(\mathrm{x}) > t \\ (2t^{-1}v(\mathrm{x}) - 1)^2 & \text{if} \quad 0.5t < v(\mathrm{x}) < t \\ 0 & \text{if} \quad v(\mathrm{x}) < 0.5t \end{cases} \tag{5.18}$$

5.2.2.2 Relative Minimised Targets

For relative minimised target requirements, such as mass or size, the opposite is true. A specific target exists, below which improvement is unnecessary, whilst a 50% increase above this target could be considered unacceptable. This function is shown graphically in Figure 5-3.

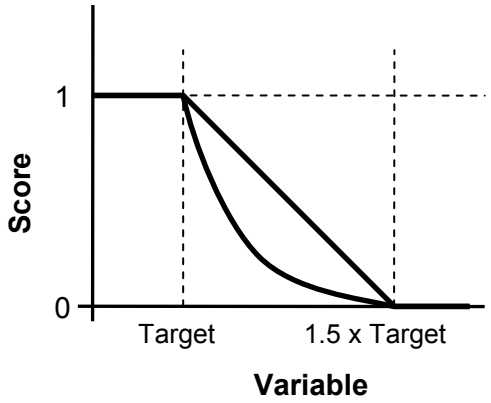


Figure 5-3: Scoring functions for relative minimised target variables

For $t < v(\mathbf{x}) < 1.5t$ the linear function shown in this figure is:

$$s(\mathrm{x}) = -2t^{-1}v(\mathrm{x}) + 3 \tag{5.19}$$

Whilst the curved function would simply be:

$$s(\mathrm{x}) = (-2t^{-1}v(\mathrm{x}) + 3)^2 \tag{5.20}$$

Again, this only applies where, so hence for relative maximised target, the complete scoring function would be:

$$s(\mathrm{x}) = \begin{cases} 1 & if \quad v(\mathrm{x}) < t \\ (-2t^{-1}v(\mathrm{x}) + 3)^2 & if \quad t < v(\mathrm{x}) < 1.5t \\ 0 & if \quad v(\mathrm{x}) > 1.5t \end{cases} \tag{5.21}$$

5.2.2.3 Absolute Maximised Targets

Both overall efficiency and power variation are unique in that their range is quite well defined. Hence unique scoring functions for these technical characteristic have to be created since the values they can take are more limited, and smaller changes are more important. For example an efficiency of 70% would clearly be unacceptable, however would still score relatively highly if the existing scoring function was used. In Chapter 4, a simple power function was used to exaggerate the difference between different calculated efficiencies; however this can no longer be used since it would violate several of the requirements of the scoring functions discussed

previously. Perhaps the biggest problem is that the range of scores yielded from this type of function would not be consistent with the other functions.

Unlike the other characteristics, the power variation coefficient and the overall efficiency are absolute values and hence do not require specified targets. Instead the target values can simply be the maximum theoretical obtainable values. Using the efficiency calculations and the optimisation process described in Chapter 4, together with the more realistic cage traction coefficient of 0.06, the maximum theoretical efficiency of the CP-CVT is very close to 95%, and hence this is the target than will be used. Next it is necessary to determine what should be considered an unacceptable value of efficiency. This can be derived simply based on the average efficiency of 1,000 combinations of dimensions, which was found to be approximately 88%, hence anything below this could be considered unacceptable. The function that describes this behaviour is shown in Figure 5-4:

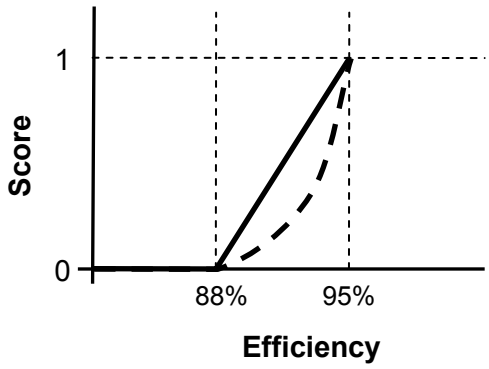


Figure 5-4: Scoring functions for efficiency

As before, this function can be squared to differentiate between different solutions, however now there are only two possibilities, since the efficiency cannot be higher than the target value:

$$s_{\eta}(\mathrm{x}) = \begin{cases} 0 & if \quad \eta(\mathrm{x}) < 88\% \\ (\frac{1}{0.07}\eta(\mathrm{x}) - \frac{0.88}{0.07})^2 & if \quad 88\% < \eta(\mathrm{x}) < 95\% \end{cases} \tag{5.22}$$

The power variation coefficient likewise has a very definable range, which varies from 0, indicating no correlation, to 1, which indicates perfect correlation between the between the horizontal displacement of the toroidal disc and the change in force required to maintain a constant torque input, implying constant power is achievable using only a simple spring. Unlike efficiency however, what is acceptable or not is somewhat ambiguous, hence for simplicity the scoring function will simply be the power variation coefficient squared:

$$s_{pv}(\mathrm{x}) = \left(p.v(\mathrm{x})\right)^2 \tag{5.23}$$

5.2.3 Relative Weighting

In order to assess the overall fitness of any set of dimensions (**x**), it is necessary to determine the relative importance of each technical requirement. It cannot necessarily be assumed that each desired parameter carries equal weighting; furthermore weightings cannot simply be assigned arbitrarily since they vary depending on each application of the CP-CVT. One of the aims of this optimisation method is to reduce the subjectivity of the process, and this includes the assigning of weighting factors. One method of reducing subjectivity is the use of a decision matrix that utilises the specific customer demands. The premise is that by using a 'relative-importance matrix', the importance of each customer demand can be determined automatically, instead of using an arbitrarily assigned score. This concept is based around Pugh's concept selection method (Pugh, 1981), described in earlier chapters.

Using a House of Quality, the importance of these customer demands can then be translated into specific technical requirements, which are weighted automatically based on their influence on each of the customer demands. This process was used in Chapter 3 to determine the weightings of relative importance for each technical demand, for three different vehicle applications, as shown in Figure 5-5.

		Application			**Technical Characteristics**						
		Family Car	**Performance Car**	**Passenger Vehicle**	Device efficiency	Transmission ratio range	Torque capacity	Mass	Length	Diameter	Output power variation
Customer Demands	Low noise	16	12	30	1	3					3
	Reduced fuel consumption	36	26	36	9	9		1			3
	Ease of implementation	16	14	10		3			9	9	
	Performance	14	42	36	3		9	9			1
	Drivability (fast response)	24	36	10	3	3		3			9
	Reliability/Durability	14	18	36	3		9				
	Low cost	30	14	12				3	3	3	
	Importance for Family Cars (Relative)				0.21	0.2	0.1	0.13	0.1	0.1	0.16
	Importance for Performance Cars (Relative)				0.19	0.15	0.19	0.19	0.06	0.06	0.17
	Importance for Passenger Vehicles (Relative)				0.22	0.17	0.24	0.16	0.05	0.05	0.12

Figure 5-5: Complete House of Quality for transmission design with specific applications

5.2.4 Target Value Setting

The previous few sections on evaluation criteria are perhaps best demonstrated using an example. First, it is necessary to determine target values for each of the technical requirements based on competition. This is a particularly difficult task since the technical characteristics of competitors' designs are not always available in literature. Furthermore, there is a large discrepancy between different designs depending on their intended application.

Kluger and Long (1999) indicate that the typical efficiencies of manual, automatic, belt-type and toroidal CVTs, are 96.2%, 85.3%, 84.6% and 91% respectively. However more recently, Yamamoto, Matsuda and Hibi (2001), showed a half-toroidal traction drive to have an efficiency of 95%. This is largely irrelevant however since transmission efficiency is design-specific and hence a target has already been imposed (95%). The ratio range likewise varies considerably with application. A typical manual transmission for a family car may have a ratio range of 0.81-3.59, giving a normalised ratio of 4.43, whilst a bus or a large truck with 12-18 gears requires a wider range to account for a wider variety of loads. Similarly the size and mass of the transmission used in a family car is relatively small compared to a truck or coach. The setting of targets is made more complicated by the fact that mass, length and size are indicative values rather than absolute.

Because of the complications of determining competitor technical characteristics, it is easier to look at the range each technical characteristic typically takes for the CP-CVT. These are summarised in Table 18.

Table 18: Typical technical characteristics of 1000 random solutions

Technical Characteristic	*Average*	*Maximum*	*Minimum*
Overall Efficiency (%)	87.6%	94.8%	45.5%
Normalised Transmission Ratio	3.16	14.13	1.07
Maximum Torque In (*Nm*)	125.9	512.7	2.5
Indicative Mass (*kg*)	30.8	226.4	1.5
Indicative Length (*m*)	0.138	0.916	0.029
Indicative Diameter (*m*)	0.144	0.256	0.062
Power Variation Coefficient	0.91	1.00	0.00

Looking at these ranges, a number of targets are proposed for each application, as shown in Table 19.

Table 19: Application dependent technical characteristic target values

Technical Characteristic	*Family Car Target*	*Performance Car Target*	*Bus or Coach Target*
Overall Efficiency (%)	95%	95%	95%
Normalised Transmission Ratio	4	5	7
Maximum Torque In (*Nm*)	180	250	350
Indicative Mass (*kg*)	40	45	60
Indicative Length (*m*)	0.15	0.15	0.2
Indicative Diameter (*m*)	0.15	0.18	0.2
Power Variation Coefficient	1.00	1.00	1.00

5.2.5 Summary of Solution Evaluation

Now that a set of targets has been established, the relative scores of each criterion can be determined. These will be based on the dimensional set already used in the example calculations shown previously (β = 45°; γ = 75°; R = 0.06m; R_1 = 0.09m; r_0 = 0.12m). The results of these calculations are shown earlier in Table 17. Given that the efficiency target is the same for each of the three applications, this will be the simplest score to calculate. Looking at Equation 5.22, the calculated efficiency for the dimensions shown is 93.3%, which is above the minimum acceptable efficiency (88%), hence:

$$s_\eta(\mathrm{x}) = (\tfrac{1}{0.07}\eta(\mathrm{x}) - \tfrac{0.88}{0.07})^2 = (\tfrac{1}{0.07} \times (0.933) - \tfrac{0.88}{0.07})^2 = 0.757^2 = 0.57$$

From the HoQ shown in Figure 5-5, the relative weightings for efficiency for family cars, performance cars, and large passenger vehicles are 0.21, 0.19 and 0.22 respectively, hence the weighted efficiency scores for each application is simply:

Family Car: $0.57 \times 0.21 = 0.12$

Performance Car: $0.57 \times 0.19 = 0.11$

Large Passenger Vehicle: $0.57 \times 0.22 = 0.13$

The calculation of the fitness score for torque capacity is slightly more complicated since each application has a different target. The calculated maximum torque input for these dimensions is 181Nm, whilst the targets for family cars, performance cars, and large passenger vehicles are 180Nm, 250Nm, and 350Nm respectively. Hence:

Family Car: $T(\mathrm{x}) > t_T$, hence $s_T(\mathrm{x}) = 1$

Performance Car: $0.5t_T < T(\mathrm{x}) < t_T$, hence $s_T(\mathrm{x}) = (2t_T^{-1}T(\mathrm{x}) - 1)^2 = 0.20$

Large Passenger Vehicle: $T(\mathrm{x}) < t_T$, hence $s_T(\mathrm{x}) = 0$

From the HoQ the relative weightings for torque capacity are 0.10, 0.19 and 0.24 respectively, hence the weighted torque capacity score for each application is:

Family Car: $1 \times 0.10 = 0.1$

Performance Car: $0.20 \times 0.19 = 0.04$

Large Passenger Vehicle: $0 \times 0.24 = 0$

Using the same principles, the remainder of the scores and weighted scores can be calculated for every technical characteristic for each application. The weighted scores are then summed to give an overall score for this particular dimensions vector set for each application. All of this information can be presented in a specially designed chart. A description of this is shown in Figure 5-6, whilst the completed sheet is shown in Figure 5-7.

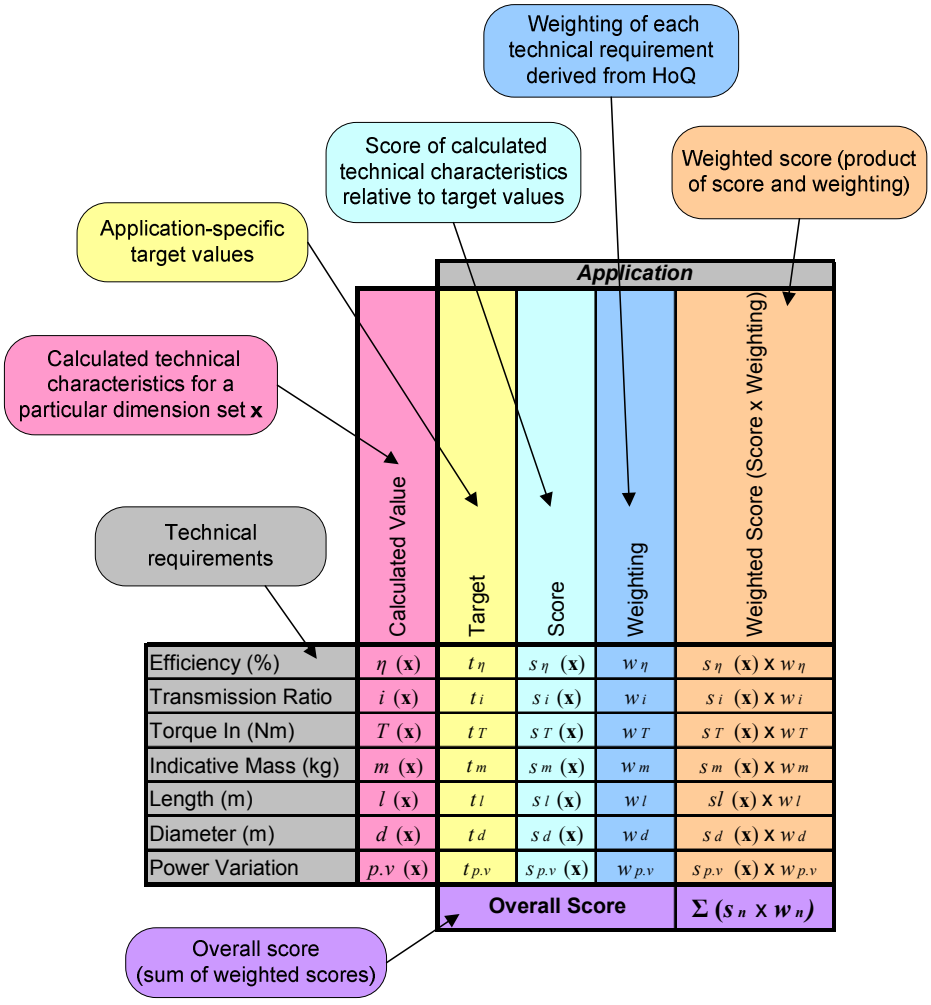


Figure 5-6: Illustrated example of the determination of overall fitness score

		Family Car				Performance Car				Bus or Coach			
	Calculated Value	Target	Score	Weighting	Weighted Score (Score x Weighting)	Target	Score	Weighting	Weighted Score (Score x Weighting)	Target	Score	Weighting	Weighted Score (Score x Weighting)
Efficiency (%)	93.3%	95%	0.57	0.21	0.12	95%	0.57	0.19	0.11	95%	0.57	0.22	0.13
Transmission Ratio	2.63	4	0.10	0.20	0.02	5	0.00	0.15	0.00	10	0.00	0.17	0.00
Torque In (Nm)	181	180	1.00	0.10	0.10	250	0.20	0.19	0.04	350	0.00	0.24	0.00
Indicative Mass (kg)	56.7	40	0.03	0.13	0.00	45	0.23	0.19	0.04	60	1.00	0.16	0.16
Length (m)	0.228	0.15	0.00	0.10	0.00	0.15	0.00	0.06	0.00	0.2	0.52	0.05	0.02
Diameter (m)	0.18	0.15	0.36	0.10	0.03	0.18	1.00	0.06	0.06	0.2	1.00	0.05	0.05
Power Variation	0.382	1	0.15	0.16	0.02	1	0.15	0.17	0.02	1	0.15	0.12	0.02
		Overall Score			**0.30**	**Overall Score**			**0.27**	**Overall Score**			**0.37**

Figure 5-7: Completed example of overall, application-specific fitness scores

The overall fitness scores for each application are relatively low, indicating that this particular set of dimensions would not be appropriate for any of the applications shown. For a family car, the main problems are the mass and size of this design, whilst the relatively small transmission ratio range and power variation are also causes for concern. For a performance car, none of the offered characteristics are close to the targets set, with the exception of efficiency. Likewise for a large passenger vehicle (bus or coach), with the exception of mass, which is more flexible in this application, the remaining characteristics are poor.

The set of dimensions used were based purely on those proposed by Glovnea and Cretu (2005), which have not been systematically optimised. Using two different optimisation processes the overall fitness score can be improved significantly.

5.3 Methodology 1: Fuzzy Swarm Optimisation

The fuzzy swarm method developed in Chapter 4 provides a starting point for multi-criteria optimisation, since it requires very little modification. Whilst originally the algorithm used efficiency as a single-scoring criterion, it must now be adapted to use the scoring functions described instead. This will provide useful information regarding the nature of the search space and where the best solutions for each application may lie, and indeed if there is any major difference in the dimensions offered for different applications.

5.3.1 Results

In Chapter 4 it was determined that each overall score must be raised by a power factor to differentiate between higher scoring solutions, hence the same principle has been applied here as well. Using a power factor of 10, a spread of 5%, and a new-point probability function value of 2 (explained in Chapter 4), the following results were obtained:

Table 20: Multi-criteria fuzzy-Swarm algorithm results

		β [°]	γ [°]	R [m]	R_1 [m]	r_0 [m]	*Overall Score*
Family Car	Average Value	14.4	48.9	0.0533	0.0628	0.0885	0.83
	Variation	1.8	5.4	0.007	0.010	0.006	0.02
Sports Car	Average Value	10.6	56.2	0.0534	0.0656	0.1044	0.82
	Variation	1.3	6.5	0.0056	0.0041	0.0061	0.01
Bus or Coach	Average Value	6.2	63.4	0.0577	0.0696	0.1301	0.79
	Variation	1.0	6.8	0.009	0.011	0.009	0.01

The average values shown are taken from running the optimisation algorithm ten times, whilst the variations show the maximum deviation found from these averages in absolute terms, rather than percentages.

The higher variations in the angle γ compared to β is confirmed on a distribution plot of the conical disc angles shown in Figure 5-8. The figure shows which areas the algorithm chose to focus on for the application of a family car. The elongation of the darker areas in the direction of the γ-axis indicates that the value of β is more restricted in terms of good solutions, whilst the value of γ is more flexible.

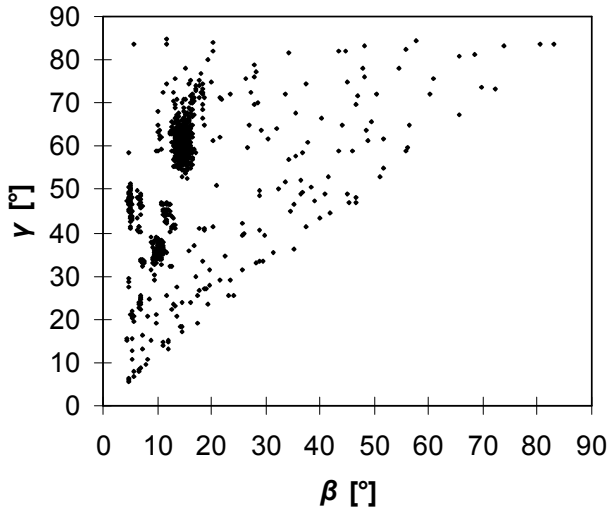


Figure 5-8: Distribution of conical disc angles for multi-criteria fuzzy-swarm optimisation

The dimensions shown say nothing about how well each solution fulfils the technical requirement targets. To determine this each application must be analysed individually. It is interesting to note however, that the initial predications made (summarised in Table 8) regarding the nature of the dimensions are relatively accurate.

5.3.1.1 Family Car

In order to determine the effectiveness of the algorithm, Figure 5-9 shows the completed overall fitness function chart for the best set of results found, using the averaged values shown in Table 20.

	Family Car				
	Calculated Value	Target	Score	Weighting	Weighted Score (Score x Weighting)
Efficiency (%)	91.5%	95%	0.25	0.21	0.05
Transmission Ratio	4.7352	4	1.00	0.20	0.20
Torque In (Nm)	180.35	180	1.00	0.10	0.10
Indicative Mass (kg)	34.828	40	1.00	0.13	0.13
Length (m)	0.14	0.15	1.00	0.10	0.10
Diameter (m)	0.15	0.15	0.99	0.10	0.10
Power Variation	0.9832	1	0.97	0.16	0.15
		Overall Score			0.84

Figure 5-9: Fuzzy-swarm algorithm-derived fitness scores for family car

This figure shows the best set of technical characteristics determined by the fuzzy-swarm algorithm. The set of dimensions derived (shown in Table 20) are capable of perfectly fulfilling four of the seven technical requirements (ratio range, torque capacity, mass and length), meeting and exceeding the targets, whilst the diameter and power variation are very close to the targets set. The only characteristic that it is not able to fulfil is the overall efficiency, which is far lower than the theoretically attainable value.

5.3.1.2 Performance Car

A similar table can be produced using the dimensions and associated technical characteristics for the intended application of a performance car, as shown in Figure 5-10.

	Performance Car				
	Calculated Value	Target	Score	Weighting	Weighted Score (Score x Weighting)
Efficiency (%)	90.0%	95%	0.08	0.19	0.02
Transmission Ratio	6.635	5	1.00	0.15	0.15
Torque In (Nm)	324.38	250	1.00	0.19	0.19
Indicative Mass (kg)	43.091	45	1.00	0.19	0.19
Length (m)	0.1285	0.15	1.00	0.06	0.06
Diameter (m)	0.17	0.18	1.00	0.06	0.06
Power Variation	0.9876	1	0.98	0.17	0.16
		Overall Score			0.83

Figure 5-10: Fuzzy-swarm algorithm-derived fitness scores for performance car

As with the family car, the algorithm yielded very strong results for the majority of the technical requirements, meeting or exceeding all except for efficiency and power variation. The power variation is extremely close to the target value, indicating that it is generally acceptable; however as with the family car, the efficiency is far lower than the desired value, yielding an absolute score of only 0.08.

5.3.1.3 Bus or Coach

The overall fitness score for the application of the CVT to a large passenger vehicle (bus or coach) is a little lower than for the other applications, as shown in Figure 5-11. Once again the majority of technical requirements have been met, with the exception of efficiency. The

increased perceived importance of efficiency in this application (indicated by its higher relative weighting) is the only reason the overall fitness score is lower.

	Bus or Coach				
	Calculated Value	Target	Score	Weighting	Weighted Score (Score x Weighting)
Efficiency (%)	89.4%	95%	0.04	0.22	0.01
Transmission Ratio	10.086	10	1.00	0.17	0.17
Torque In (Nm)	378.54	350	1.00	0.24	0.24
Indicative Mass (kg)	45.286	60	1.00	0.16	0.16
Length (m)	0.1468	0.2	1.00	0.05	0.05
Diameter (m)	0.19	0.2	1.00	0.05	0.05
Power Variation	0.9902	1	0.98	0.12	0.12
		Overall Score			0.79

Figure 5-11: Fuzzy-swarm algorithm-derived fitness scores for bus or coach

5.3.2 Discussion

The relatively low fulfilment of the efficiency criteria in all of the applications could either be because the relative weighting of efficiency is not high enough, or dimensions that yield a higher overall efficiency yield overall poor technical properties. The fact that the efficiency already has one of the highest relative weightings for all applications indicates that the latter is more likely. This was confirmed by artificially increasing the relative-weighting of efficiency, which simply resulted in lower overall fitness scores and only a minor improvement in efficiency.

Further sensitivity analysis of each input dimension implies that the algorithm functions well, as shown in Figure 5-12. In these figures the solid black lines indicate the overall score, whilst the blue symbols indicate efficiency. The vertical dotted is the algorithm's suggested dimension. For the majority of the input dimensions the dotted line is on, or close to, the highest possible score, and this rarely coincides with the maximum efficiency value. This implies that according to the methodology used, including the use of a house of quality and Pugh's concept matrix, it is better sacrifice efficiency in favour of having better overall characteristics in order to satisfy customer demands.

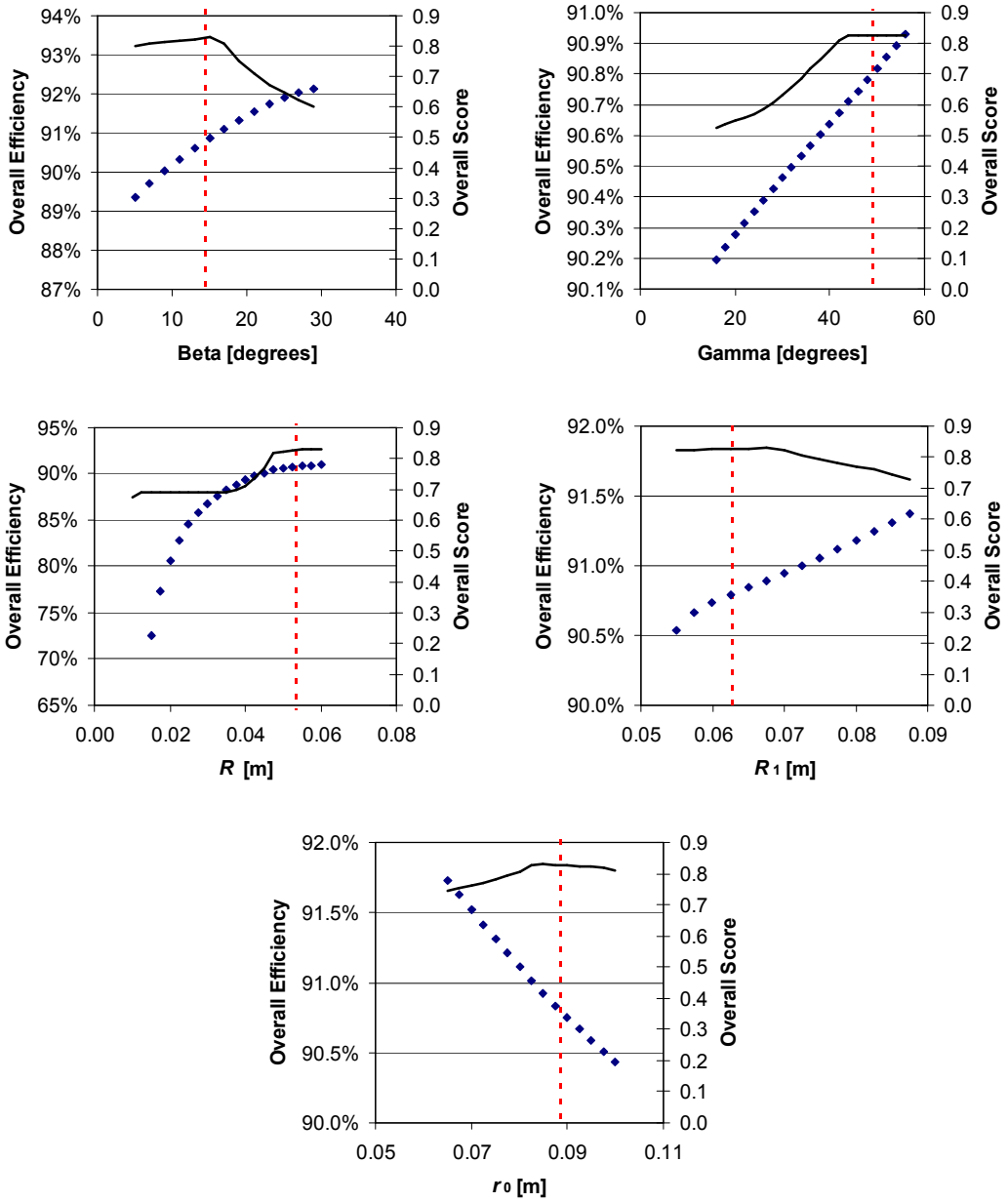


Figure 5-12: Influence of input dimensions on overall efficiency and overall fitness score

One of the problems with the fuzzy-swarm technique is that it only yield regions of interest rather than specific or consistent values, which is highlighted by the variation in the suggest dimensions shown in Table 20. Although the algorithm is capable of quickly finding high scoring areas, the chance of it actually finding the precise point is based on luck, and is analogous to throwing a dart at random into a five-dimensional dartboard and hoping to hit a specific target. In reality, the 'best' possible solution can never be guaranteed unless all possible solutions are evaluated (brute-force search), however certain algorithm techniques, such as evolutionary or genetic algorithms aim to increase the chances of finding the best solution by using the principle of natural selection.

5.4 Methodology 2: Genetic Algorithm

5.4.1 The Basic Genetic Algorithm

Genetic algorithms (GA) are computational methods that are based on the principles of evolution and natural selection. The basic premise is that solutions can be evolved to create better and stronger solutions through the concept of 'survival of the fittest', eventually culminating in the strongest solution possible (Whitley, 2001). This process involves 3 main stages:

1. Creation of the initial population
2. Chromosome selection and intermediate generation creation
3. The creation of the next generation

5.4.1.1 Creation of the Initial Population

Initially a number of solutions (chromosomes) are generated at random to form the initial population (generation-zero). These solutions, which can number several thousand, are typically distributed randomly around the entire search space, although it can be preferable to deliberately 'seed' a number of known good solutions.

5.4.1.2 Chromosome Selection

The initial (or previous) generation's chromosomes are evaluated and scored based on their ability to fulfil certain criteria (fitness score). Typically a roulette-style selection method is used to select the stronger solutions from this generation, which are carried forward to fill an intermediate generation. Certain solutions maybe selected twice, whilst others may not be selected at all, although it is generally favourable to give every solution at least a small probability of being selected.

5.4.1.3 Breeding and Creation of the Next Generation

Solutions from the intermediate generation are then selected to 'breed' the next generation. This involves selecting 'parent' solutions (typically a pair), which are crossed-over to produce a 'child', which contains some of the characteristics of each parent. This results in a new generation of solutions, which are theoretically different and stronger than the previous generation. This process is then repeated until termination. A mutation operator can also be applied to randomly alter a particular solution, introducing an element of diversity to a generation. The purpose of this is to avoid local maxima by preventing generation of solutions from becoming too similar to each other, which can slow or even halt subsequent evolution.

5.4.1.4 Termination

A genetic algorithm can be designed to run for a predetermined length of time (a specified number of generations for example), or a halting criteria can be introduced. This can involve running the algorithm until a target is reached, or until each successive generation is no longer yielding progressively improved solutions.

5.4.2 Implementation

One might reasonably question the advantage of using a genetic algorithm over more traditional search or optimisation algorithms. The most common argument put forward in defence of genetic algorithms is their ability to robustly and efficiently search through many different solutions (Whitley, 2001). The multi-criteria optimisation of the CP-CVT can be viewed as a search problem in which there is a 5-dimensional hyper-cube of possible solutions, where each dimension represents one of the critical dimensions of the CP-CVT and the aim is to find the *best* solution. According to Wright et al. (2005), a genetic algorithm is essentially able to search many different planes within this hyper-cube simultaneously, significantly reducing the computational time required. Each new generation samples many solutions from this hyper-cube at the same time, evaluates them then directs the search accordingly. This type of search algorithm is thus ideally suited to multi-criteria dimensional optimisation problem.

Previously, the fuzzy-swarm optimisation algorithm was simplistic enough to be implemented in Microsoft excel, however the vast number of variables, conversions and calculations required for genetic algorithms requires the use of a more sophisticated programming language. The methodology described herein utilises Microsoft Visual Studio, and specifically the Visual Basic language.

5.4.2.1 Determining Chromosome Length

Traditionally in genetic algorithms each chromosome consists of a series of binary bits of a predetermined length, which together form a single value or combination of values, whilst evolutionary algorithms take the raw value without conversion to a binary number. There remains a lot of debate in literature about which approach is better (Whitley, 2001; Holland, 1992; and Biles, 1994). The key components of the CP-CVT (conical input and output disc, toroidal disc and intermediary element) can be described through five input dimensions, two angular (β and γ) and three linear (R, R_1, r_0), as described previously. Assuming for now that the use of binary is a better approach, the resolution of each variable is hence determined by the length of the binary string used for each input. These strings are then concatenated to form a complete chromosome. The use of binary strings means that the number of discrete values an

input value can take is limited. Assuming that it is desirable to not have any null bit string combinations (every possible string has a defined equivalent dimensional value), then the resolution of each dimension is determined by bit length and the imposed limits. The lower and upper limits for the angles of the input and output discs can be set to 0° and 90°, whilst the limits for the radial dimensions can be set to 0m and 0.25m. Table 21 shows the effect of bit length on the resolution of the angle and linear dimensions.

Table 21: String length effect on input dimension resolution

String length	*Number of discrete values*	*Resolution of Angle Dimensions*	*Resolution of length dimensions 0.25m limit*	*Total Number of Combinations Available*
5	32	2.813°	7.813mm	3.36E+07
6	64	1.406°	3.906mm	1.07E+09
7	128	0.703°	1.953mm	3.44E+10
8	256	0.352°	0.977mm	1.10E+12
9	512	0.176°	0.488mm	3.52E+13
10	1024	0.088°	0.244mm	1.13E+15

If it is desirable to have an accuracy of at least 1mm, then 8 bits are required for each dimension. The total gene length is thus 40 bits long (8 bits for each of the 5 dimensions). This means that each dimension is discrete and can take any one of 256 possible values.

In order to populate the initial gene pool, each dimension is randomly assigned an integer value from 0 to 255, which corresponds to a specific discrete value between the upper and lower limits. In order to demonstrate this, let us take a random integer value of 25. This corresponds to 11001 in binary (00011001 when padded to ensure consistent string length). For an angular dimension this would correspond to an angle of:

$$25/255 \times 90 = 8.82^{\circ}$$

Whilst for a linear dimensions, this would correspond to:

$$25/255 \times 0.25 = 0.0245\text{m}$$

5.4.2.2 Populating the Gene Pool

Populating the initial gene pool involves randomly assigning values to each of the five input dimensions. It is desirable at this stage to guarantee each set of dimensions is at least feasible by ensuring it fulfils a number of geometrical constraints. This can be implemented in a number of different ways:

1. Assign values to each dimension in a random order, limiting the range each dimension can take based on dimensions already assigned. *This would be the most efficient solution in terms of processing times, however would require a more complicated algorithm simply to populate the initial pool.*
2. Treat physical constraints as a criterion for the fitness score of each chromosome and assign a zero score to those that fail. *This would be the simplest solution but would lead a high proportion of chromosomes being excluded immediately.*
3. Apply a physical constraint test after each chromosome has been created and recreate the chromosome if it fails the test. *The advantage of this solution is that no limits are placed upon the values that each dimension can take, whilst failed solutions can be continuously replaced until a 'good' solution is found.*

Whilst method 2 might initially seem like the most logical approach, initial tests with random dimension sets indicated that approximately 95% of randomly assigned combinations of dimensions fail a geometrical-constraint test. Hence if this approach was used then a large number of erroneous combinations would have to be ignored at each evolution, severely limiting the number of good chromosomes that can be carried forward. Further experimentation indicated that if two feasible sets of dimensions are combined then the chance of the yielding another feasible set of dimensions is a lot higher (approximately 70-80%). Given these findings, it is very important to ensure that the initial gene pool is populated with physically feasible chromosomes, hence method number 3 is the most appropriate.

5.4.2.3 Physical Constraints

Previously very simple dimensional constraints were implemented using only four rules. However, as was shown in Chapter 4, these constraints are not always sufficient. From Figure 5-13, there are six constraints that must be imposed:

- $2R < \sqrt{2[(R_1 - R)\sin 57]^2}$: *Assuming four ball elements there must be a gap between the ball elements when they are closest to the shaft*
- $\beta < \gamma$: *To maintain force balance as discussed previously*
- $R < R_1$: *The toroidal radii must be greater than the radii of the ball elements*
- $x_1 < x_2$: T*he edges of the conical and toroidal input discs should not contact* $y_1 < y_2$: *The tip of the toroidal disc should not interfere with the movement of the conical output disc*
- $y_3 >$ shaft radius: *When the ball elements are closest to the shaft, their edge should not be in contact with it*

The values of x_1, x_2, x_3, and y_1, y_2 can be determined from the coordinates of each point derived previously in Section 4.3.5.

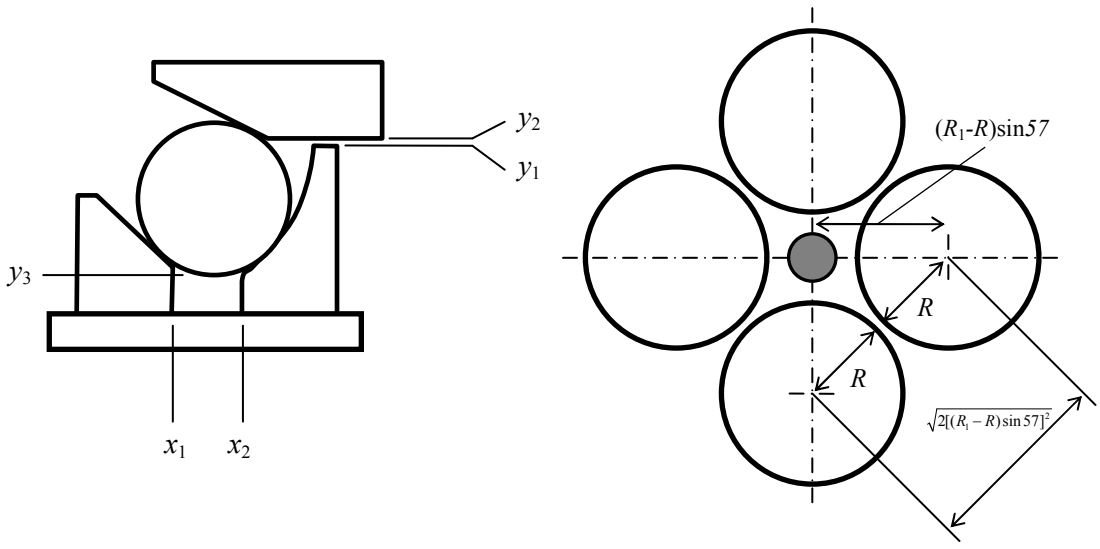


Figure 5-13: Physical dimensional constraints

The implementation of these dimensional constraints, using method 2 shown on the previous page can be accomplished within a computer program very simply by adding a Boolean function and a Do-Loop, as shown in Appendix A.5

5.4.2.4 Distribution of Input Dimensions

Given that each dimension is assigned randomly, one might reasonably expect that the probability distribution of each dimension should be approximately even across all possible values. However the physical constraints prevent this from occurring. Looking again at conical disc angles, angle β must be less than γ, and hence will tend towards smaller values, whilst γ will be more likely to take larger values, as shown in Figure 5-14.

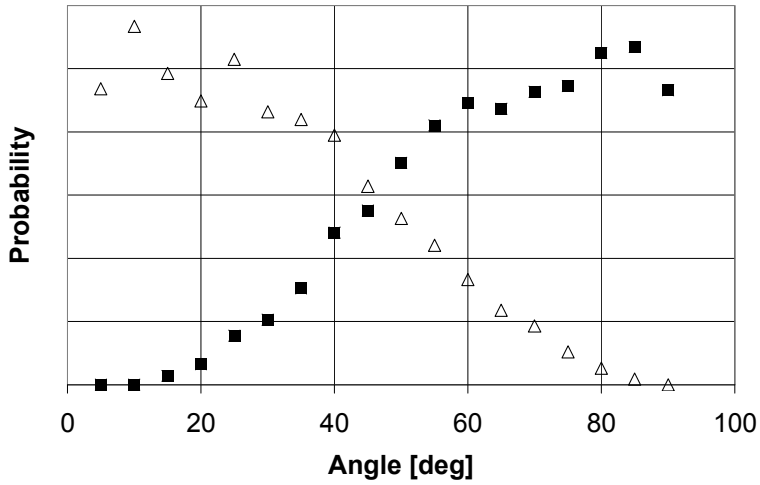


Figure 5-14: Probability distribution of input angles β (Δ) and γ (■)

Similar and more complicated restraints were placed upon the linear dimensions, which leads to the distribution pattern shown in Figure 5-15.

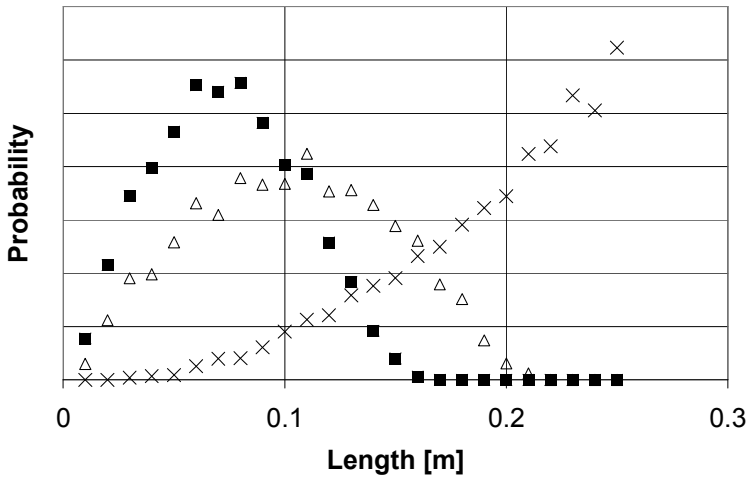
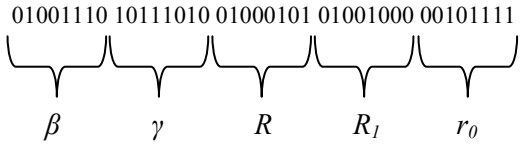


Figure 5-15: Probability distribution of input length dimensions R (■), R_1 (Δ) and r_0 (X)

5.4.2.5 Implementation of Cross-Over

For genetic algorithms that use bit-strings, breeding pairs of chromosomes involves determining a random crossover point (or points), then taking part of one chromosome up to this crossover point and combining it with the remaining bits from the other chromosome. In some GAs cross-over does not always occur when populating a subsequent generation from an intermediate generation. Occasionally a parent chromosome can be in the same generation as offspring in order to ensure certain strong solutions are not disregarded through reproduction with weaker chromosomes. The probability of parents being combined is known as the cross-over rate. In this modified genetic algorithm, rather than using a fixed rate, an intelligent cross-over routine is implemented. After each cross-over, an immediate feasibility test is conducted, if the new chromosomes fail the constraint test, the chromosomes are disregarded and the parent chromosome is instead carried forward unaltered to the next generation.

One of the problems with the way the complete chromosome has been created is that each of the chromosomes consists of the five input variables in a fixed order, as shown below.



If a single cross-over point is used then adjacent dimensions, such as β and γ have a higher probability of staying together, whilst those further away have a lower probability. One method of overcoming this is to have more than one cross-over point. By introducing a second cross-over point, the string can be viewed as a loop (De Jong, 1975), somewhat eliminating this problem. In essence, this leads to a greater number of solutions being tested, since there is more chance of separating and testing 'good' dimensions.

5.4.2.6 Implementation of Mutation

Potentially after several evolutions, certain bits will be fixed to either 0 or 1 since the inverted bit is essentially extinct and no longer part of the generation. Genetic mutation attempts to overcome this by randomly changing a single bit value of a chromosome in order to ensure that no particular bit is ever prematurely fixed to a particular value (De Jong, 1975). The probability of this occurring is deliberately set very low to ensure that it doesn't disrupt the natural evolution of solutions. Rather than using mutation, the GA utilised here incorporates random seeding. Instead of randomly changing a single bit value, occasionally a completely new chromosome is added to a generation replacing an existing one. This has the advantage of ensuring feasible solutions are continuously seeded into the gene pool.

5.4.3 Results

The results produced from the GA for the application of a family car are shown in Table 22. A comparison of these results to the results produced from the fuzzy swarm algorithm, and the technical characteristics that these dimensions yield are shown later. Initially however it is necessary to ensure that the algorithm is behaving as expected. To achieve this, there are two main observational tools that have been employed: the evolution of the dimensions, and the spread of the results.

Table 22: Results produced from genetic algorithm for the application of a family car

Parameter	*Value*
β	19.7°
γ	56.8°
R	0.0610m
R_1	0.0686m
r_0	0.0891m

5.4.3.1 Evolution of Dimensions

It was stated previously that ideally in a genetic algorithm each generation should produce a stronger set of chromosomes. Indeed in some GAs the lack of any improvement between generations is used as a halting criterion. However continuous subsequent improvement was not found to occur for this particular optimisation problem. Instead it was found that typically the majority of improvement occurred over the first 100-200 evolutions, whilst beyond this improvement still occurred, but at a significantly decreased rate, as shown in Figure 5-16, which shows how the maximum overall score changed over 1000 evolutions.

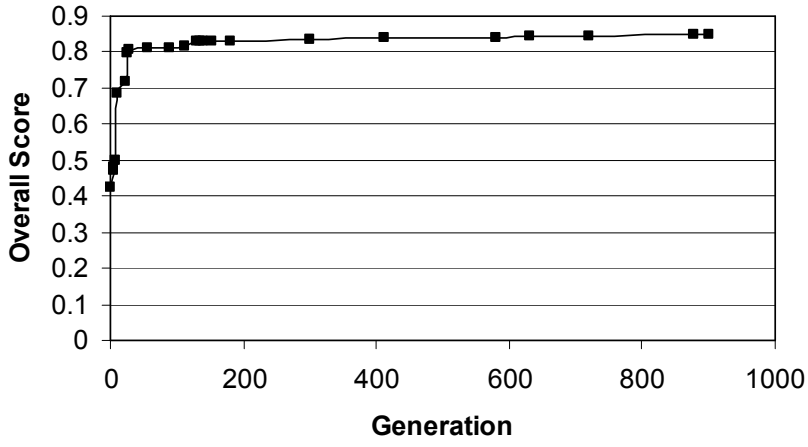


Figure 5-16: Evolution of overall score

This figure shows that improvements, albeit minor can still occur even after over 100 evolutions have occurred without improvement. This is also reflected in the evolution of the dimensions, shown in Figure 5-16, which shows the value of each dimension that yields the fitness scores shown in the previous figure.

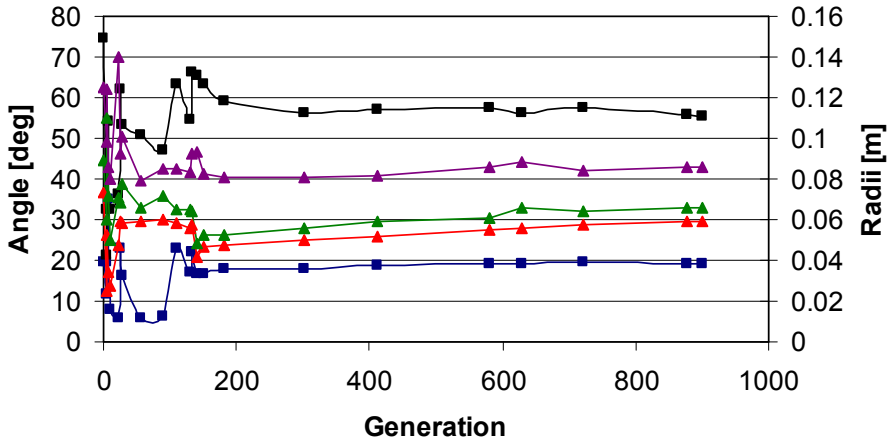


Figure 5-17: Evolution of angle dimensions (■), β (blue) γ (black) and radii dimensions (▲), R (red), R_1 (green) and r_0 (purple)

As before there is a lot of initial variation in the dimension values, reflecting the large initial improvements. These variations tend to decrease as the generations become more homogenous and less diverse. The reason there isn't a continuous increase across every generation is possibly because of the nature of the problem, with each dimension having no specific relationship to one another and hence an improvement that results from the combination of chromosomes is somewhat due to luck.

From Figure 5-16 and Figure 5-17 it is clear that the majority of evolution and improvements happen over the first 100-200 generations, whilst after this only minimal improvements occur. This implies that approximate solutions can be obtained very quickly by running the algorithm with large initial sample sizes and fewer evolutionary steps, however exact solutions require far more evolutions.

5.4.3.2 Number of Cross-Over Points

One criterion for measuring the effectiveness of an optimisation algorithm is to determine the spread of solutions offered from repeatedly running the algorithm with the same input criteria. Theoretically an exact algorithm should continuously yield the same results. With a complex

problem, such as dimensional optimisation, it is perhaps unreasonable to expect a perfect correlation, since the algorithm is designed to offer the best output parameters, which might conceivably be fulfilled with more than one combination of input dimensions. Furthermore, the difference in overall score between small changes in the input dimensions is marginal, as shown in Figure 5-16 and Figure 5-17.

Perhaps the best way of showing the overall spread of the solutions derived from genetic algorithm repetitions is through the use of a probability distribution function. This function identifies the probability of a value (x) falling within a particular interval of the mean (μ). The standard equation for probability distribution is:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Where σ and μ are the standard deviation and mean of the given data respectively.

An exact algorithm should produce a very low deviation from the mean, which is represented in probability distribution function plot as a very narrow, sharp curve. The wider the distribution curve, the more varied the given results are, and hence the less consistent the algorithm.

Within the genetic algorithm there are a number of parameters that can be controlled, such as the mutation rate, the crossover rate, the initial population size and the number of generations. It was found through experimentation with the GA that stronger results could be obtained when using a relatively large initial population size (1000), and a fixed number of evolutions (1000), whilst the effect of using either a 1-point or a 2-point crossover is shown in Figure 5-18.

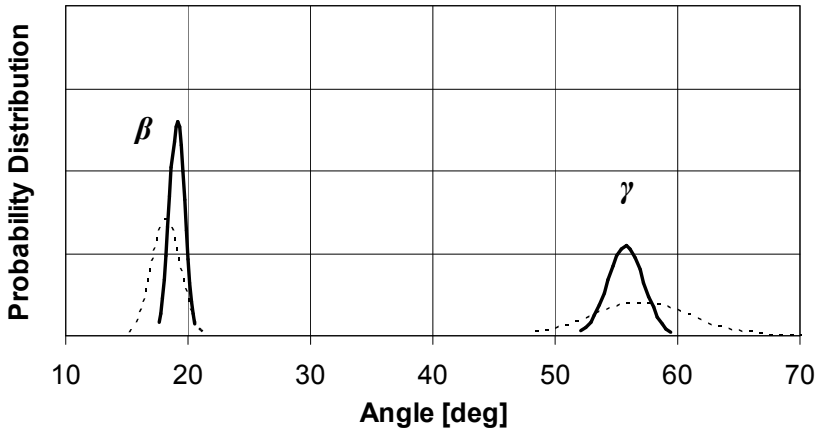


Figure 5-18: Probability distribution of β and γ using 1-point (dotted) and 2-point (solid) crossover

To produce this figure the resulting values of β and γ were found across 100 identical algorithm runs using either 1-point or 2-point crossover, utilising the intelligent crossover rate described earlier. Given that this comparison uses an identical number of evolutions (1000), it is entirely possible that given enough time both 1-point and 2-point crossover rates would yield identical solutions, however in terms of computational efficiency it is clear that the use of 2-point crossover yields superior and more accurate results. Interestingly, the same findings were not reflected in the distribution of the radial dimensions, as shown in Figure 5-19.

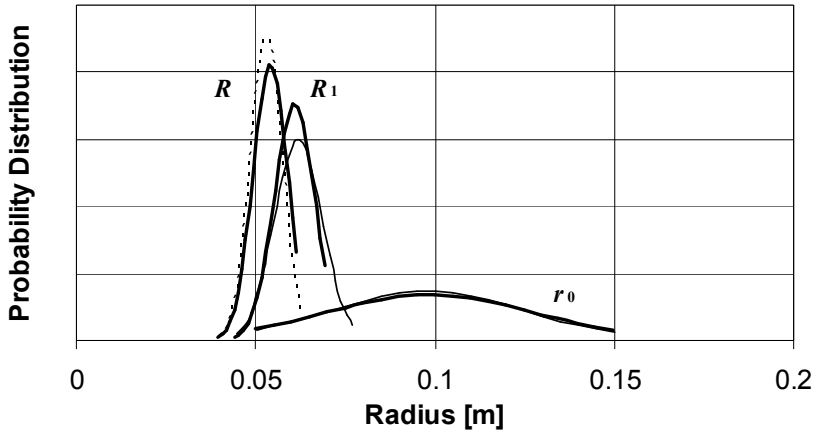


Figure 5-19: Probability distribution of R, R_1 and r_0 using 1-point (dotted) and 2-point (solid) crossover

Whilst the advantage of using 2-point crossover is reflected in the distribution of R_1, the opposite appears to be true for the value of R. This maybe because in the chromosome layout R is the centrally concatenated dimension and hence there is more chance of disturbing this value and hence disrupting its natural evolution. The large variation of r_0 confirms that there is more flexibility in its value in relation to the overall characteristics of the CVT, as was found earlier (Figure 5-12).

5.4.3.3 Mutation/Seeding Rate

As discussed previously, rather than using a mutation operator, this specific GA uses random seeding instead. This is designed to ensure that generations do not become too homogonous by deliberately seeding a random solution that will at least pass the dimensional constraint test. As with mutation, the probability of seeding (seeding rate) has to be very low (approximately 1-2%). If it is set too high it begins to disturb the natural evolution resulting in much wider spread of distributed results across repeated GA runs. Seeding is very important however, as is demonstrated in Figure 5-20. This graph shows the evolution of each dimension only when an improvement in the overall score occurs. This particular algorithm deliberately used a small

initial sample size (increasing the likelihood of homogeneity amongst generations) and was run across 10,000 generations. The parts on the graph where the lines start to show only minimal changes show premature convergence, whilst sudden changes, especially later on, show where a new seed has been introduced, stimulating further evolution.

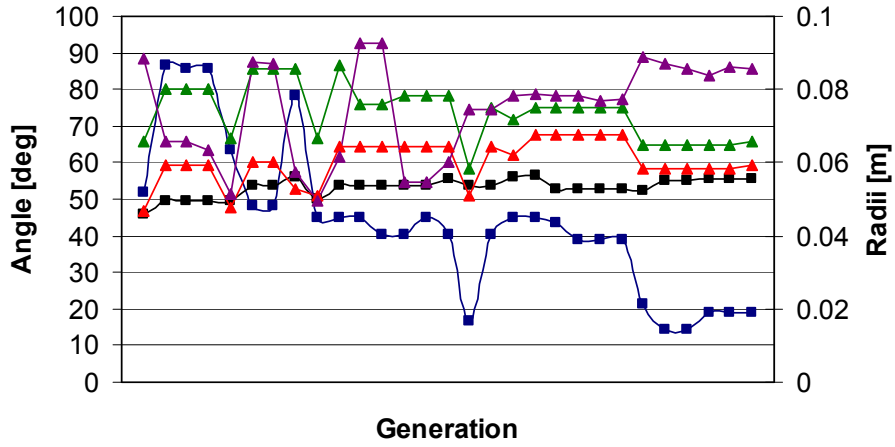


Figure 5-20: Evolution of angle dimensions (■), β (blue) γ (black) and radii dimensions (▲), R (red), R_1 (green) and r_0 (purple) when improvement occurs

5.5 Comparison of Genetic and Fuzzy Swarm Algorithm Results

5.5.1 Summary

Based on the findings of the previous sections, the most efficient operating parameters for this particular GA have been found. Hence the following results were produced using a seeding rate of 2%, 2-point intelligent cross-over, with an initial sample size of 1000 chromosomes evolved over 1000 generations. Table 23 shows the results produced by the GA compared with those produce by the fuzzy-swarm technique.

Table 23: Comparison of dimensional results from GA and fuzzy-swarm techniques

	Algorithm Used	β [°]	γ [°]	R [m]	R_1 [m]	r_0 [m]	*Overall Score*
Family Car	Fuzzy-Swarm	14.4	48.9	0.0533	0.0628	0.0885	0.83
	Genetic Algorithm	19.7	56.8	0.0610	0.0686	0.0891	0.85
Sports Car	Fuzzy-Swarm	10.6	56.2	0.0534	0.0656	0.1044	0.82
	Genetic Algorithm	13.7	54.6	0.0660	0.0739	0.0739	0.84
Bus or Coach	Fuzzy-Swarm	6.2	63.4	0.0577	0.0696	0.1301	0.79
	Genetic Algorithm	6.8	57.9	0.0726	0.0806	0.124	0.79

The 'exact' dimensions produced by the GA are slightly different from the results of the fuzzy-swarm optimisation. However the additional computational time required for the GA optimisation has only marginally improved the overall score for each application, with the most significant improvement being shown in the family and sports applications. The actual improvements yielded from the use of a GA are shown in Figure 5-21.

	Family Car			Performance Car			Bus or Coach		
	Target	Fuzzy Swarm Calculated Value	GA Calculated Value	Target	Fuzzy Swarm Calculated Value	GA Calculated Value	Target	Fuzzy Swarm Calculated Value	GA Calculated Value
Efficiency (%)	95%	90.8%	91.9%	95%	89.9%	91.0%	95%	88.9%	89.7%
Transmission Ratio	4	4.51	4.05	5	6.39	5.21	10	10.91	10.04
Torque In (Nm)	180	198.6	180.4	250	275.9	250.3	350	413.5	351.1
Indicative Mass (kg)	40	24.4	31.1	45	27.6	44.1	60	38.2	59.9
Length (m)	0.15	0.12	0.13	0.15	0.13	0.14	0.2	0.14	0.16
Diameter (m)	0.15	0.14	0.14	0.18	0.16	0.17	0.2	0.19	0.20
Power Variation	1	1.00	0.97	1	0.99	0.99	1	0.99	0.99
Overall Score		**0.827**	**0.848**		**0.826**	**0.843**		**0.782**	**0.791**
Improvement		**2.59%**			**2.09%**			**1.19%**	

Figure 5-21: Technical characteristic improvements produced from the use of a GA

From this figure it is clear that the GA was able to improve the overall score for each application by improving the efficiency of the design. This has invariably come at a cost of reductions in other areas such as the transmission ratio range, and mass. Despite this, each technical characteristic has still met the target; hence the overall score has improved. This implies that the results produced by the original fuzzy-swarm method were 'over-engineered' by going beyond what was required. This is perhaps best illustrated by the maximum torque input. The fuzzy-swarm algorithm suggested dimensions that allowed a torque input far higher than what was required, whilst the refinements suggested by the GA reduced the maximum torque input to almost precisely the target value. This allowed an increase in efficiency of around 1%, improving the overall score by a similar amount. The point is further demonstrated in Table 24, which shows the percentage fulfilment relative to the target of each technical requirement produced by both optimisation methods for the application of a family car.

Table 24: Fulfilment of technical requirements for each algorithm for a family vehicle

Technical Characteristic	*Fuzzy Swarm Fulfilment*	*GA Fulfilment*
Overall Efficiency	96%	97%
Normalised Transmission Ratio	113%	101%
Maximum Torque In	110%	100%
Indicative Mass	164%	129%
Length	125%	116%
Diameter	106%	104%
Power Variation Coefficient	99%	97%

Values less than 100% indicate that the target has not been met, whilst values over this indicate over-engineering. The high percentages shown for the dimensions produced from the fuzzy swarm algorithm indicate a high degree of over engineering, and hence inefficient design. Although this table is specific to a family car, similar results were found for the other applications. The results shown here and the increases in overall score indicate that despite the increased complexity and computational time associated with using a GA, it does produce superior results.

The implementation of the dimensions shown is discussed in more detail in later sections.

5.5.2 Modifications to Fuzzy Swarm Optimisation

Although the fuzzy-swarm methodology produced inexact results (more diversity in the offered dimensions) and lower overall fitness score, it remains a very simple methodology to implement and operate. Currently the algorithm uses only a thousand particles to assess the entire search space, hence requiring only a thousand evaluations of the technical characteristics. By comparison the genetic algorithm requires 1,000,000 evaluations (a thousand initial chromosomes, and a thousand evolutions). By increasing the number of particles used in the fuzzy swarm algorithm, the consistency of the results and the overall score can be improved substantially. One method of doing this is to use random-restart seeding. This method, which is commonly used in hill-climbing, restarts the algorithm at certain points. The criteria for restarting can either be fixed (a pre-determined number of steps), or function based, such as when the algorithm results start to show little improvement. The fuzzy-swarm algorithm described previously was adapted to use the latter criteria. On restart the algorithm clears all knowledge of existing points, and begins again with the initial points being chosen from the highest scoring points of the previous run. Further particles are then assigned with a significantly reduced new-point probability, decreased spread, and increased scoring factor.

This adaptation to the fuzzy-swarm optimisation algorithm essentially explores the search space iteratively, initially searching the entire search space, and then incrementally focusing on smaller search regions each time the algorithm restarts. Using these modifications the fuzzy-swarm technique produced near-identical results to the genetic algorithm, and required far fewer evaluations of the technical characteristics.

5.5.3 Modifications to Genetic Algorithm

The results discussed previously imply that better solutions can be found by having larger initial sample sizes with fewer evolutions. The optimum sample size and number of evolutions have yet to be determined, but generally depend on finding a balance between accuracy, robustness, and computational time. Further experimentation with the algorithm indicated that a further improvement can be made by using smaller sample sizes, fewer generations, and running the algorithm repeatedly. The best solution can then be chosen from the solutions given by each run, or an average can be taken of all the results. Additionally, each restart can be set to include good results already found. This modification, which is essentially a form of parallel evolution, avoids some of the common problems associated with standard genetic algorithms such as premature convergence since by taking many different samples from the initial gene pool diversity is still maintained in the overall population (Whitley, 2001).

5.6 Conclusions

This chapter has presented two successful methods of multi-criteria optimisation for the dimensions of the CP-CVT. Target values and relative weightings were derived for three specific automotive applications through the use of a customised House of Quality and general quality function deployment techniques. Specialist scoring functions were also created and utilised in both optimisation algorithms to assess the overall fitness of any particular design.

Initially, the previously designed fuzzy-swarm optimisation technique was used to determine the approximate values of each application-specific dimension. The results indicated that according the design approach used, customer satisfaction would be improved by sacrificing transmission efficiency in favour of improvements in other technical characteristics. Furthermore the results showed that provided particular component dimensions are used, the CP-CVT is capable of fulfilling more than 80% of the technical requirements, and by extension, 80% of customer demands. The largest disappointment for customers is a small reduction in the desired efficiency of the system, which still remains superior to automatic transmissions.

These findings were confirmed through the use of a genetic algorithm. The genetic algorithm was capable of producing more consistent results, with a higher overall fitness score. This was achieved by reducing several of the technical requirements to values closer to their targets thus allowing an improvement in the transmission efficiency. In order to produce the results, the canonical genetic algorithm was modified for this specific optimisation problem. These modifications included the use of seeding rather than mutation and the use of an intelligent cross-over rate that always attempts to combine strong chromosomes, unless the result yields a null set of dimensions that fail a physical constraint test. The results shown demonstrate the usefulness of these modifications by quickly yielding good sets of dimensions that can be immediately used in the development of the CP-CVT for automotive purposes. Furthermore, now the algorithm has been created and implemented it is a straightforward task to adapt it for different sets of targets and weightings. Hence a genetic algorithm, once designed, provides a very efficient design tool for multi-criteria optimisation and rapidly changing targets and weightings. In general they are considered extremely useful because of their ability to quickly evaluate non-linear, interdependent relationships. In essence they provide an efficient method of searching a multi-dimensional array of potential solutions, where the alternative approach would be to systematically test every possible solution, which may be impractical and time consuming.

Brute-force searching of the localised areas indicated by the results indicated that no appreciable further improvement could be made to the overall score for the targets and weighting used;

hence the results produced by the genetic algorithm shown in Table 23 and Figure 5-21 can be considered exact. Suggestions were made for possible modifications that could be applied to both optimisation techniques. The addition of random restarting to the fuzzy swarm optimisation technique allowed it to produce near-identical results to those produced by the genetic algorithm, which have been shown to be exact. This was accomplished with far few evaluations of the fitness criteria and hence required far less computational time. The modifications suggested for the genetic algorithm (parallel evolution), cannot significantly reduce the computational time required, which is a big criticism of the method, but do allow the validity of the results to be assessed by monitoring the homogeneity of the suggested dimensions across a number of algorithm repetitions.

CHAPTER 6: VEHICULAR SIMULATION

Parts of this chapter have been presented by the author at the 2009 JSME International Conference on Motion and Power Transmissions and published in the associated proceedings.

6.1 Introduction

6.1.1 Chapter Summary

Now that the CP-CVT has been dimensionally optimised and shown to be capable of fulfilling customer requirements, the next stage is to determine how the transmission system would behave and respond in a vehicular environment. Financially it is not viable at this stage of development to implement the CP-CVT in a real vehicle, and so a simulated model must be used instead. Hence this chapter shows the development of a simulation, based on first principles, that monitors the movement and velocity of each of the key components and determines how these affect the acceleration and behaviour of the vehicle. Each aspect of the vehicle, such as the engine behaviour, and tire friction is appropriately modelled using fundamental mathematical equations. By resisting the temptation to use a pre-existing software package, the resulting simulation is fully customisable and easy to use, thus making it a potentially useful tool for further vehicle-specific optimisation later on.

Part of the purpose of this simulation is to determine if it is plausible to control the CP-CVT automatically with a fixed loading system. Thus far it has been assumed that a spring-type loading system would be sufficient. This chapter attempts to determine whether that assumption is valid.

Whilst other parameters can be optimised through basic calculations, a real time model is perhaps the only method of optimising the dimensions of the CP-CVT to offer the best possible response time and hence drivability.

6.1.2 Background

Continuously variable transmission technology currently offers one of the greatest potentials to reduce automotive emissions. The race for the next widely implementable CVT design is thus very lucrative. It is generally accepted that toroidal traction-drives currently have the most potential of the various CVT designs since they can offer improved wear-resistance over belt-type CVTs through the use of a traction fluid, which prevents the metallic surfaces of the key components from coming into direct contact. The torque-controlled CP-CVT described previously is one such toroidal traction-drive design that offers novel solutions to several drawbacks currently shown in other designs. These drawbacks include complex and costly loading systems, and convoluted control mechanisms required to adjust the transmission ratio and synchronise all contacting parts.

One problem often cited as the reason that CVTs have not been more widely accepted is the relatively poor response time, which is partially due to the complex control system required in current CVT designs. Response time and hence drivability is a difficult concept to definitively describe. If response time is considered the time it takes for the CVT to respond to a change in a driver's demanded ratio, then drivability could be considered to be the ability of the CVT to effectively *predict* driver ratio demand. This is extremely complicated however, as can be demonstrated by a simple example: Let us assume that a driver is accelerating, and has now reached the desired speed. The driver now reduces the throttle pedal position, thus indicating that they wish reduce acceleration. There are now two possible drive scenarios:

1. *The driver wishes to remain at the current speed, for example when cruising on a highway.* The transmission ratio should thus decrease to allow the engine speed to reduce, generally improving fuel efficiency.
2. *The driver wishes to slow down gently, for example after overtaking another vehicle or when travelling down a gradient.* The transmission ratio should thus increase slightly to employ engine 'braking', reducing the need to use the normal brakes, which waste energy.

This simple situation highlights the complexity of determining the ideal transmission ratio for any situation. However before discussing complex situations such as this, it is necessary to determine what the CP-CVT would do without any external control. Hence one of the aims of this simulation is to determine how response time and drivability can be improved, with or without external control, and preferably without adversely affecting fuel consumption.

6.2 Methodology

6.2.1 Simulation Type

There are several existing techniques that could be employed for the simulation of vehicular motion, including the use of existing software packages (McManus and Anderson, 2005), or more advanced fundamental simulations that employ a finite-element, lumped-mass approach (Dutta-Roy, 2004). Existing software packages are not ideally suited to specific or custom transmission designs, and allow little control of specific aspects of the simulation. The finite-element, lumped-mass approach is more widely used for fundamental analysis, since it allows the determination of stability, vibration and noise of each component as a function of source harmonics and internal damping. In order to demonstrate this method, Figure 6-1 shows a simple gear pair modelled as lumped masses (Dutta-Roy, 2004)

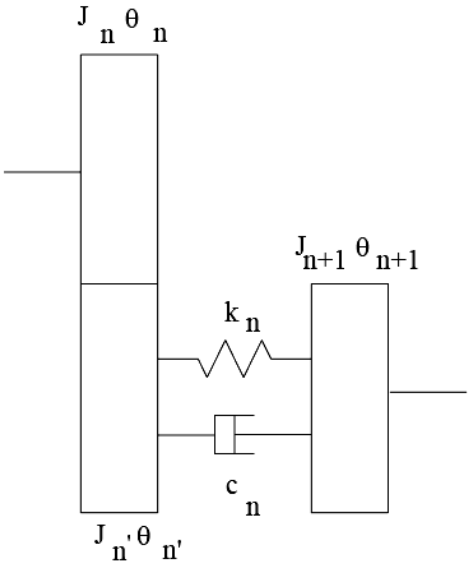


Figure 6-1: Simple, lumped-mass gear paid

The dynamics of each component are formulated using finite elements with independent co-ordinates, each with individual damping coefficients, lumped moments of inertia and either constant or variable stiffness. These aspects are then assembled in a global system matrix, where parameters such as free and forced vibration can be calculated using the eigenvalues and eigenvectors of the system matrix. The size of the global system matrix grows exponentially with the number of components and hence this approach is generally suitable for smaller, simpler components, such as gear pairs. Whilst it has been successfully applied to a complete CVT system (Dutta-Roy, 2004), the complexity of the process warranted an entire thesis. Furthermore, because each component is essentially calculated independently at each time-interval based on adjacent component damping and stiffness forces, the time-interval must be

very small to ensure system stability. If it is set too high then the system will be very unstable, as shown in Figure 6-2 (Zhang and Dutta-Roy, 2004).

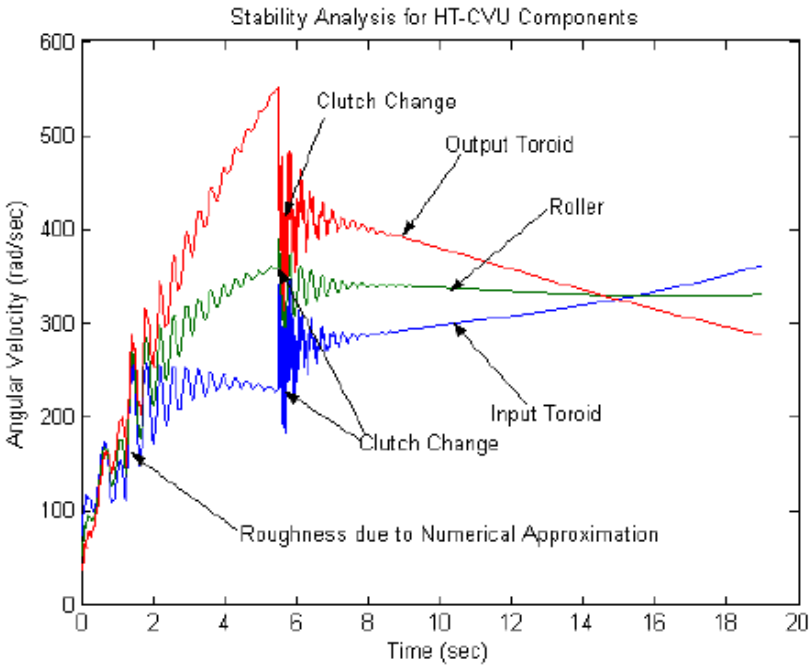


Figure 6-2: System stability due to insufficient system damping and large time interval

Conversely, if the time-interval is too small, then the computational time and storage required quickly become unacceptable. A better and possibly simpler solution is to use exact numerical modelling. This approach uses a set of models to simulate all mechanical components in the transmission, both at a macro-scale (engine, vehicle, road), and at a micro-scale (CVT components), allowing a progressive and focused analysis of the system (Fuchs, Hasuda and James, 2000). The precise nature of the sub-models can be based either on experimental data, numerical simulation, or precise equations, meaning each sub-model can usually represent real behaviour reasonably accurately.

6.2.2 Macro-Level Model

On a macro scale, a vehicle fitted with the CP-CVT consists of three primary components, as shown in Figure 6-3:

1. The vehicle's engine
2. The CP-CVT and its associated components
3. The interaction between the tire and road (and hence implicitly movement of vehicle)

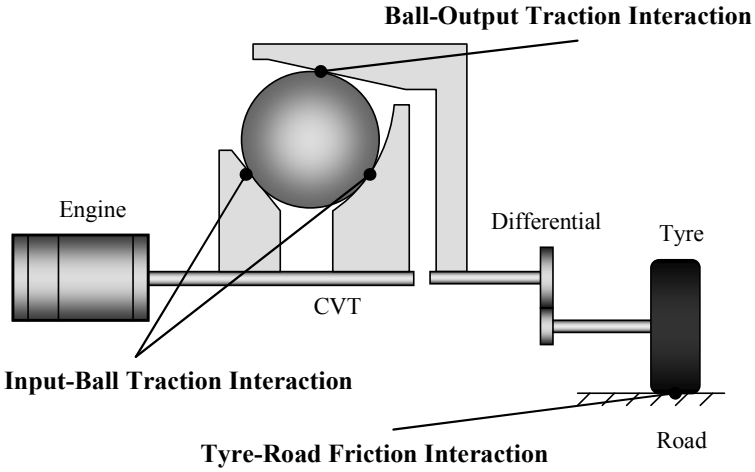


Figure 6-3: Sub-model interaction

The engine sub-model is simply based on a numerical approximation of empirically obtained data (in this case, engine performance data). Similarly the tire-road interaction is based on existing tire-friction models, and hence is relatively simple to implement. The main complexity of this simulation is thus in the modelling of the CP-CVT component behaviour.

6.2.3 Vehicle Engine

Real engines vary considerably depending on a wide variety of factors, such as throttle position, torque load, engine speed, and manifold pressure. Realistically it is impossible to obtain exact data for every engine that it is desirable to simulate, and hence a numerical function must be substituted that relates engine speed (ω_i) to the maximum available torque (T_e). A typical torque-speed curve obtained from empirical data is shown in Figure 6-4 (Pfiffner and Guzzella, 2001), which also shows constant power and constant efficiency lines. The data shown is for a modern 2 litre, fuel-injection, spark-ignition engine.

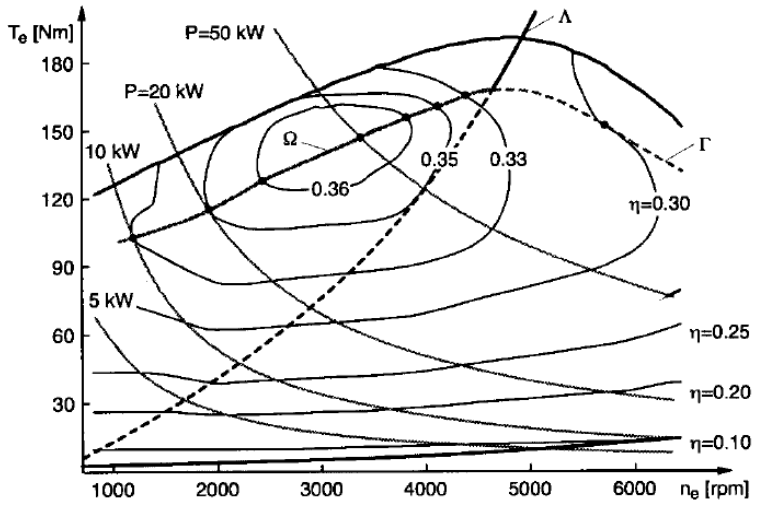


Figure 6-4: Map of a modern engine, showing iso-efficiency curves and iso-power curves.

As expected the peak-efficiency operating points occur at lower engine speeds and higher torque outputs, and furthermore, a good efficient operating curve can be achieved by running close to the maximum torque curve, which according to this figure is approximately quartic. Hence in order to simulate a range of different engine types, a simple quartic function has been derived as shown in Equation 6.1.

$$T_e = a\omega_i^4 + b\omega_i^3 + c\omega_i^2 + d\omega_i + e \tag{6.1}$$

The values of the constants *a*-*e* can be calculated automatically based on the data that is generally available for any engine:

- The maximum torque output and associated engine speed
- The torque at maximum power output and associated speed
- The turning point of the torque curve (i.e. when torque output peaks)
- The torque output when the engine speed is zero.

The full derivation of this quartic Equation is shown in Appendix A.6

It should be noted that based on this equation it is possible for there to still be torque available when the engine speed is zero. This is required for when the vehicle accelerates from rest. The alternative to this would be to use a planetary gear system that would allow the engine to rotate (idling) even when the vehicle is stationary. However, for the purposes of this simulation it is sufficient to assume that this equation instead defines the torque output of an engine combined with an automatic, slip-clutch. Implicitly this also removes the ability of the vehicle to move in reverse at this stage.

A typical torque curve produced using this equation is shown in Figure 6-5, obtained using the simulation program. This curve uses data from a 2005 Ford Mondeo 2.0L engine, which has a maximum torque output of 180Nm at 4,500rpm, and maximum power output of 107kW at 6,000rpm, as shown in the figure. The similarities between this figure and the curves shown in Figure 6-4 (which are taken from empirical data) validate the use of a quartic equation.

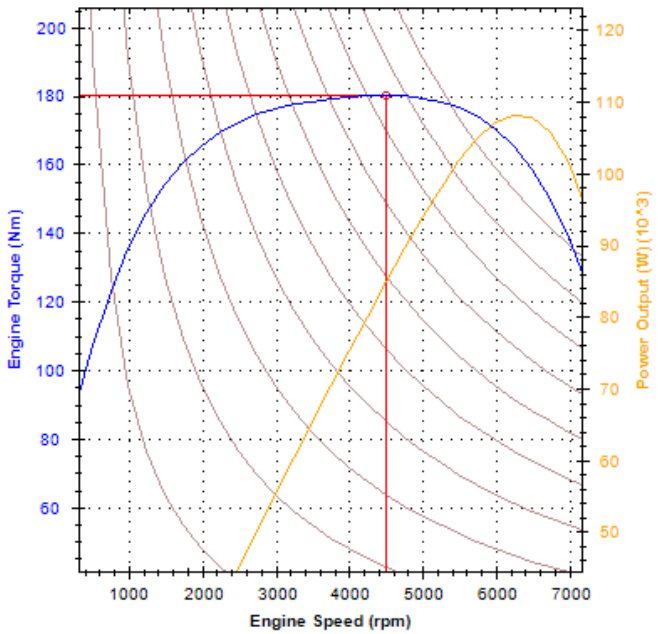


Figure 6-5: Engine torque curve produced using quartic function

The actual torque output is also a function of the throttle position. To implement this, a simple quadratic equation was developed that relates throttle position to the percentage of maximum torque delivered, as shown in Figure 6-6.

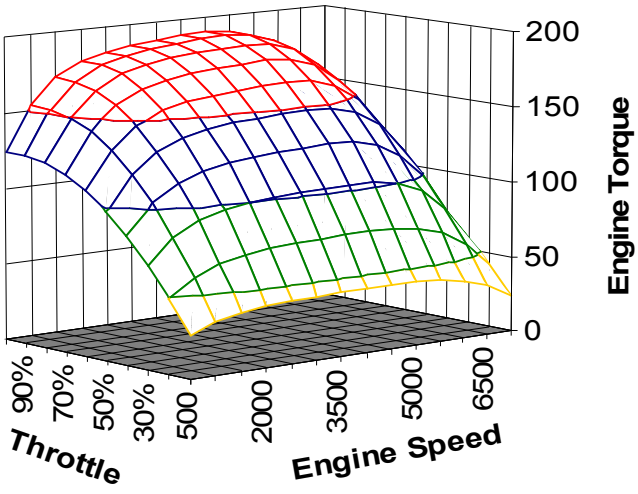


Figure 6-6: Full engine map showing influence of throttle position on engine torque

6.2.4 Tire-Road Interaction

For simplicity it is assumed here that vehicle motion is longitudinal, and hence the forces applied to the vehicle as a whole can be considered to be acting against the motion of travel. Hence the acceleration of the vehicle ($\dot{V}$) can be calculated using Equation 6.2, which assumes a level road surface.

$$\dot{V} = \frac{F_x - F_{dr}(V) - F_{rr}(V)}{m_v} \tag{6.2}$$

Where F_x is the driving force of the vehicle, and F_{dr} and F_{rr} are the forces due to drag and rolling resistance respectively. The force due to drag and rolling resistance are simply:

$$F_{dr} = \tfrac{1}{2}\rho_{air} V^2 C_D A$$
$$F_{rr} = mgC_R(V)$$

Where C_D is the coefficient of drag, A is the frontal area of the vehicle, ρ_{air} = 1.2 kg/m^3 and C_R is a function of the tire pressure, type and vehicle velocity (Short, Pont, and Huang, 2004)

The driving force can be calculated based solely on the friction between the driving tires and the road surface (μ_w), as shown in Equation 6.3.

$$F_x = \mu_w N_w \tag{6.3}$$

Where N_w is the normal force acting on the tire as a result of the vehicle weight. For accuracy this normal force must be equal the dynamic weight distribution of the vehicle. Whilst the static load distribution is simply a function of the vehicle geometry, the dynamic load distribution will change as a result of the weight distribution that occurs as the vehicle accelerates and brakes, and hence is also a function of the position of the vehicle's centre of mass.

The interaction between the tire and the road and hence the calculation of the friction coefficient is a complex phenomenon, especially during cornering or with unbalanced tire loads. Since vehicle motion is considered longitudinal, a number of friction models can be employed of which the most widely accepted is Pacejka's tire friction model (Pacejka and Sharp, 1991). This model is based on a complex curve fitted to experimental data. As shown in Chapter 7, the parameters of this model have little physical significance, and hence this model is simply an approximation based on observation. However, its simplicity and relative accuracy make it ideal for vehicular simulation. The model states that the friction coefficient is simply a function of the slip between the tire and road surface (s), as shown in Equation 6.4

$$\mu_w = c_1 \sin(c_2 \arctan(c_3 s - c_4(c_3 s - \arctan(c_3 s)))) \tag{6.4}$$

During acceleration the slip can be defined as shown in Equation 6.5.

$$s = \frac{V - \omega_w R_w}{\omega_w R_w} \tag{6.5}$$

The magnitude of tire friction available depends largely on the road surface conditions. Surfaces that are more slippery (such as loose gravel or ice) significantly reduce the tire friction coefficient. Increases in vehicle speed also decrease the tire friction coefficient, albeit to a lesser extent, however this isn't such as issue for modelling vehicular acceleration since the driving force applied to the wheel by the engine is offset by increased drag forces at higher vehicle velocities. Hence at higher velocities there is less tangential force applied directly to the road surface anyway and the actual friction coefficient is reduced. Several experimental studies have measured tire friction in various conditions to determine the Pacejka variables ($c_{1\text{-}4}$ in Equation 6.4). Typical Pacejka curves produced using these variables are shown in Figure 6-7 for dry, wet and snow conditions (Short, Pont, and Huang, 2004).

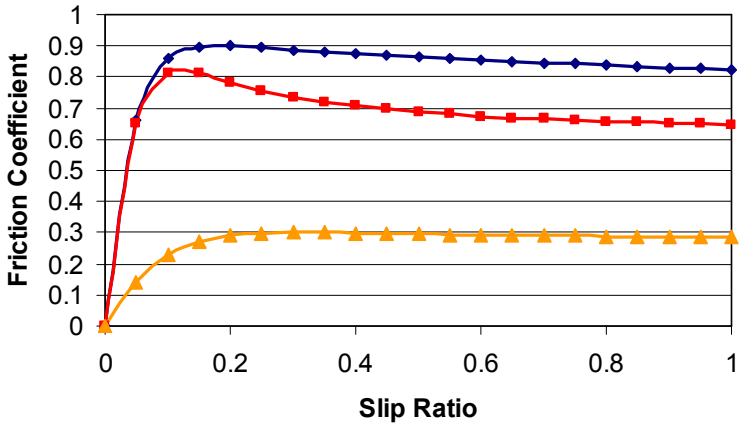


Figure 6-7: Friction coefficient in dry (blue), wet (red) and snow conditions (orange)

6.2.5 The Traction Drive

The CP-CVT traction drive is a far more complex mechanism to model, consisting of several moving parts, interactions and inertias. The effects of spin and creep have been described previously in Chapter 5, however for simplicity here only creep will be considered, since it is important to ensure that it does not regularly exceed the desired value, increasing wear. Furthermore, the cage losses have also been neglected since they do not fundamentally effect the operation and behaviour of the CP-CVT. In the following sections the meaning of β, γ, R, R_1 and r_0 are the same as those used throughout this thesis, shown again here for ease of reference.

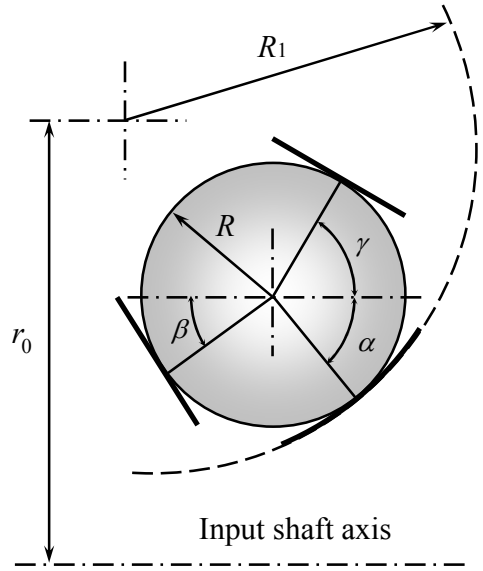


Figure 6-8: Explanation of dimensions of CP-CVT

6.2.5.1 Traction Fluid Behaviour

To accurately simulate traction, the traction fluid behaviour must be understood. To achieve this, several experimental tests were conducted on a dedicated traction fluid, with the commercial name Santotrac50. The tests were carried out on a PCS Instruments Mini-Traction Test rig (MTM). An elastohydrodynamic contact was formed between a flat disc and a ball, loaded together with a 50N load, which gave a Hertzian pressure of about 1.5GPa. The disc and the ball were driven independently at desired speeds, such that the entrainment speed was kept constant at 2.2m/s, while the slide-roll ratio varied between 0 and 0.1. The slide-roll ratio is defined as before as:

$$S = \frac{2|u_1 - u_2|}{u_1 + u_2} \tag{6.6}$$

The tests were carried out in isothermal conditions, at the expected operational temperature (80ºC). The results closely match those found earlier by Anghel, Glovnea and Spikes (2004). Based on the results, a model similar to the Pacejka friction model was devised. This provides a mathematical relationship between the traction coefficient (μ) and slide-roll ratio (S), as shown in Equation 6.7.

$$\mu = 0.105 \times \sin(1.37 \times \arctan(105 \times S - 0.99 \times (105 \times S - \arctan(105 \times S)))) \tag{6.7}$$

The correlation between the theoretical and measured traction coefficient values is shown in Figure 6-9.

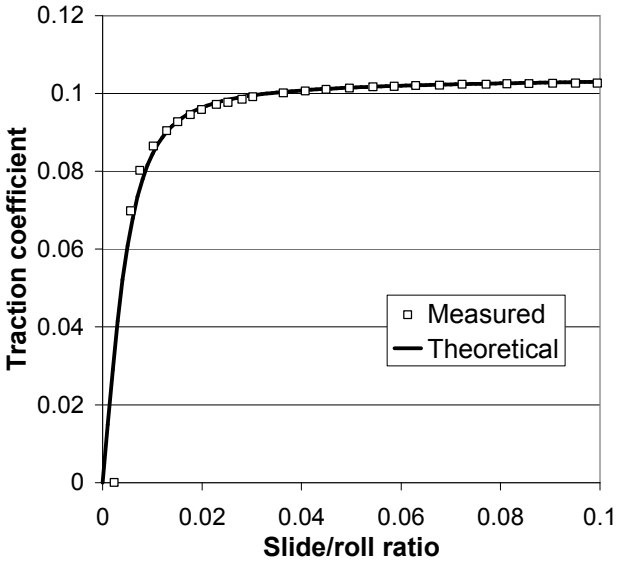


Figure 6-9: Correlation of measured and theoretical traction

In order to determine the magnitude of the traction force developed at each disc's contact point the normal forces at each contact point must be known. These forces can be calculated as a function of the torque applied to the ball screw coupling, as shown earlier in Equations 2.15-2.18.

6.2.5.2 The Traction Drive: Rotational acceleration of elements

Rather than using a stiffness-dampening model as shown previously, it is desirable here to determine an exact relationship between adjacent parts. Initially we will begin by looking at each component within the CVT individually, starting with the input shaft and input discs, which can be simplified to a rotating lumped mass as shown in Figure 6-10

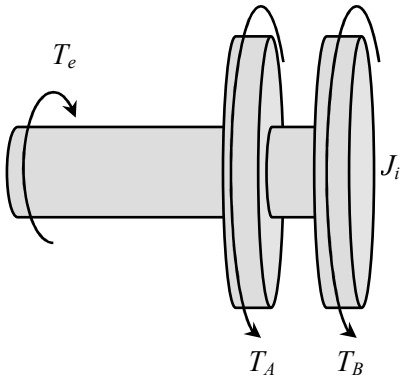


Figure 6-10: Lumped moment of inertia of input discs and shaft

Where T_e is the torque produced by the engine, and J_i is the lumped moment of inertia of the discs, input shaft and flywheel.

The equivalent torque applied to each input disc can be calculated from the tractive force applied at each contact and the distance to each point of contact from the centre of the shaft, i.e.

$$T_A = F_{tr-A} r_A \text{ and } T_B = F_{tr-B} r_B$$

These tractive forces can be determined simply from the traction coefficient at each point, and the normal force applied, hence:

$$T_A = \mu_A N_A r_A \tag{6.8}$$
$$T_B = \mu_B N_B r_B \tag{6.9}$$

The rotational acceleration of the input shaft ($\dot{\omega}_i$) is hence:

$$\dot{\omega}_i = \frac{T_e - T_A - T_B}{J_i} \tag{6.10}$$

Combining Equations 6.8-6.10 yields an expression for the rotational acceleration of the input shaft as a function of the normal forces applied to each contact:

$$\dot{\omega}_i = \frac{T_e - r_A \mu_A N_A - r_B \mu_B N_B}{J_i} \tag{6.11}$$

Similarly, the rotational acceleration of the ball elements ($\dot{\omega}_{ball}$) can be calculated using a lumped moment of inertia of all the ball elements combined (J_{ball}):

$$\dot{\omega}_{ball} = \frac{r_A \mu_A N_A + r_B \mu_B N_B - r_C \mu_C N_C}{J_{ball}} \tag{6.12}$$

The rotational acceleration of the output shaft is slightly more complex since there is an additional differential/final gear ratio (i_f) that must be considered. This mechanism can be simplified to a lumped moment of inertia as shown in Figure 6-11.

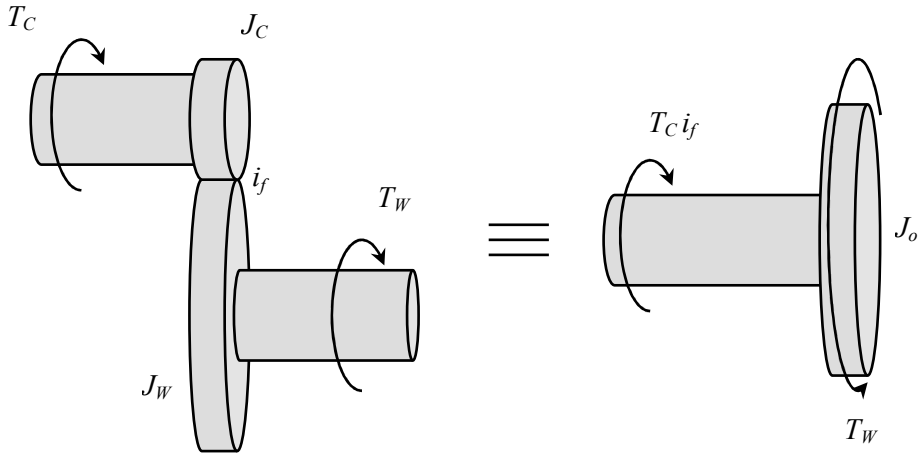


Figure 6-11: Lumped moment of inertia of output disc and wheel

Looking at the torque at the wheel, we have:

$$T_C i_f - T_w = \omega_w J_o \tag{6.13}$$

The torque applied to the wheel (T_w) can be calculated as before from the friction coefficient between the wheel and road surface:

$$T_w = \mu_w N_w R_w \tag{6.14}$$

Where N_w is the normal force due to the weight of the vehicle, applied to the driven wheels. Hence:

$$\dot{\omega}_w = \frac{(\mu_C N_C r_C) i_f - \mu_w N_w R_w}{J_o} \tag{6.15}$$

This value of J_o needs to be adapted to include the equivalent moment of inertia of the output disc applied after the final gear ratio, i.e.:

$$J_0 = J_C i_f - J_w$$

6.2.6 Complete Model

Equations 6.2, 6.3, 6.11, 6.12 and 6.15 can now be combined to provide a relationship between the acceleration of the wheels and the torque provided by the engine:

$$\dot{\omega}_w = \frac{\left(\frac{R_A r_C}{R_C r_A}\right) i_f T_e - \left(\frac{R_A r_C}{R_C r_A}\right) i_f \dot{\omega}_i J_i - \left(\frac{r_C}{R_C}\right) i_f \times \dot{\omega}_{ball} J_{ball} - R_w \left[\dot{V} m_v + F_{dr} + F_{rr}\right]}{J_o} \tag{6.16}$$

Equation 6.16 requires that the rotational acceleration of each individual component is known. The rotational acceleration of each component could simply be calculated through kinematic calculations based on the acceleration of the wheel and the relative radii of each contact point; however this would neglect the behaviour of the traction fluid. To more accurately model the traction fluid the slide-roll ratio of each contact point must be known. If it is assumed that the CVT is specifically designed to always operate within the linear region of the traction curve, then the slide to roll ratio can be simply calculated as shown in Equation 6.17.

$$\frac{2(\omega_1 r_1 - \omega_2 r_2)}{(\omega_1 r_1 + \omega_2 r_2)} = \frac{\mu}{m_{tr}} \tag{6.17}$$

Where m_{tr} is the slope of the linear region of the traction curve.

Rearranging this equation yields the relationship between the rotational speed of one component as a function of the proceeding component:

$$\omega_1 = \omega_2 \frac{r_2}{r_1} k_{tr} \tag{6.18}$$

Where:

$$k_{tr} = \left(\frac{1 + \frac{\mu}{2m_{tr}}}{1 - \frac{\mu}{2m_{tr}}} \right) \tag{6.19}$$

Hence k_{tr} indicates the loss in speed that occurs due to the contact (creep loss). k_{tr} is contact specific, referring to either the contact between the input disc and ball (k_{ib}) or the contact between the ball and the output disc (k_{ob})

If a similar expression is written for the coefficient of slip between the tire and road surfaces, then Equation 6.16 can be rewritten solely in terms of the torque produced by the engine and the rotational acceleration of the wheels.

$$\dot{\omega}_w = \frac{\left(\frac{R_a r_c}{R_c r_a} \right) i_f T_e - F_{dr} R_w - F_{rr} R_w}{\frac{R_w^2}{k_w} m_v + J_o + \left(\frac{r_c}{R_c} \right)^2 i_f^2 k_{bo} \left[J_{ball} + \left(\frac{R_a}{r_a} \right)^2 J_i k_{ib} \right]} \tag{6.20}$$

Where $k_w = f(\mu_w)$, which can be determined through reverse iteration of the Pacejka friction model (Equation 6.4).

6.2.6.1 Toroidal Disc Position

It has been determined previously that the transmission ratio of the CP-CVT (and by extension the values of $R_{A\text{-}C}$ and $r_{A\text{-}C}$) are a function of the geometry of the key components, together with the variable contact angle between the toroidal disc and the ball elements (α), as shown in Equation 6.21.

$$i = \frac{\sin(\alpha+\beta)((r_0 - (R_1 - R)\sin\alpha) + R\sin\gamma)}{(r_0 - (R_1 - R)\sin\alpha)\cos\alpha\sin\gamma - R\sin(\alpha+\beta)\sin\gamma + (r_0 - (R_1 - R)\sin\alpha)(\sin(\gamma-\beta) + \sin\alpha\cos\gamma)} \tag{6.21}$$

Furthermore, it can be shown that the position of the toroidal disc relative to its initial position (x) is also a function of α, as shown in Equation 6.22.

$$x = (R_1 - R)(1 + \sin\alpha\tan\beta - \cos\alpha) \tag{6.22}$$

Rearranging in terms of α:

$$\alpha = \arccos\left(\cos\beta\left(1 - \frac{x}{R_1 - R}\right)\right) - \beta \tag{6.23}$$

Hence the transmission ratio and the position of all of the key components can be determined from the position of the toroidal disc. This immediate position can be found from the disc's instantaneous acceleration, which is simply a function of its mass and the forces applied to it. The force applied on one side of the disc (F_A) is initially assumed to be produced by a simple spring of stiffness k, hence:

$$F_A = kx \tag{6.24}$$

The force opposing this (F_A') is created by the linear force produced by the ball-screw coupling, which can be calculated by adapting Equation 2.20 shown earlier:

$$F_A' = \frac{l}{2\pi T_{out}} \frac{(\sin\alpha\cos\gamma + \cos\alpha\tan\beta\cos\gamma)}{(\cos\alpha\sin\gamma - \cos\alpha\tan\beta\cos\gamma)} \tag{6.25}$$

where l is the ball-screw lead length.

The linear acceleration of the disc is hence obtained from:

$$\frac{d^2x}{dt^2} m_{toroidal} = F_A' - kx \tag{6.26}$$

Hence in dynamic terms, any change to torque at the output of the CP-CVT will cause a subsequent change in the force produced by the ball-screw coupling, which will in term alter the forces applied to the toroidal disc, changing its acceleration and position, forcing a shift in the transmission ratio.

6.3 Simulation Program

The model described previously has been implemented in a fully functional computer program written in Microsoft Visual Basic. This user-orientated program, which consists of approximately 4,000 lines of code, allows simple control of all aspects of the vehicular simulation as described, including the engine, tire friction properties, CVT dimensions, loading system and traction properties.

6.3.1 Engine Tab

The engine tab of the program creates the shape of the engine's torque-speed curve based on either a cubic, quartic or constant function. The user simply enters the maximum torque and power, and the engine speed at which they occur. The program will then create the curve automatically. As explained previously, the quartic function also includes a "shape" factor (Equation 6.1), which controls the initial gradient of the torque-curve, whilst still passing through the points described. A screen shot of the program in operation is shown in Figure 6-12.

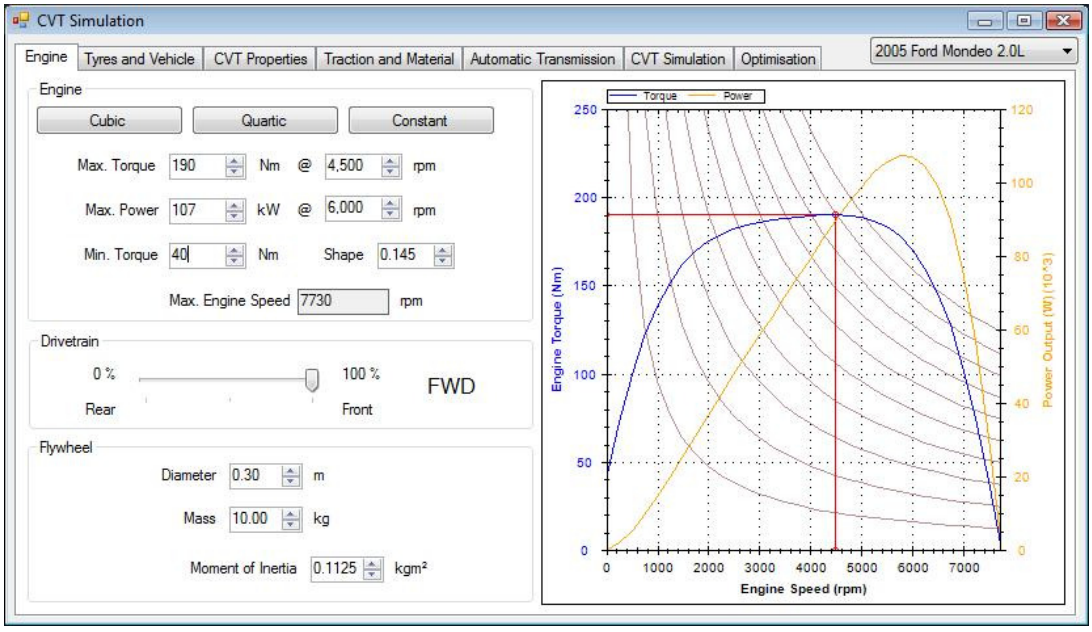


Figure 6-12: Screen-shot of engine tab of simulation program

This figure also shows how the user can control other aspects of the drive train, including the distribution of drive to either the front wheels (FWD), real wheels (RWD), or a combination of both (4WD), and the properties of the flywheel.

6.3.2 Tires and Overall Vehicle Tab

This tab primarily allows the user to control the nature of the tire friction curves that will be used based on the coefficients of the Pacejka tire friction model (Equation 6.4). In addition to being able to control each coefficient manually, a number of presets are also included for dry, wet, snow and ice conditions, as shown in Figure 6-13. The friction curve is automatically plotted for both acceleration and deceleration conditions.

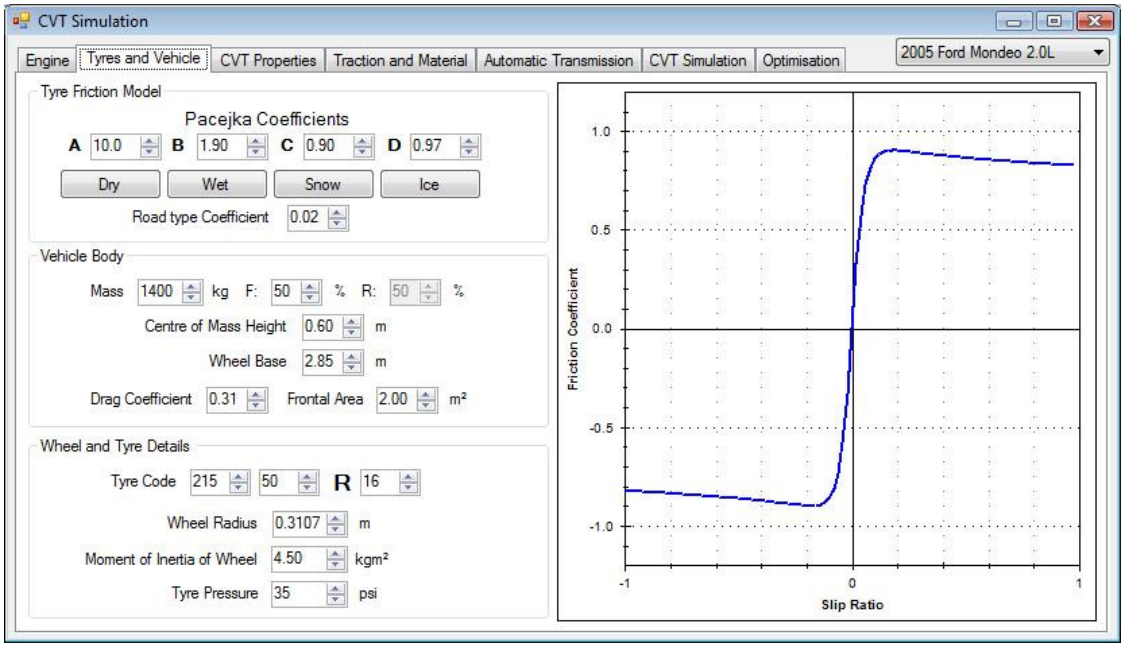


Figure 6-13: Screen-shot of tire and vehicle tab of simulation program

This tab also allows additional information to be entered regarding the vehicular body, such as the drag coefficient and frontal area, which affect drag force, and tire pressure, which affects rolling resistance. The tire radius can be calculated automatically based on the international standard tire code.

Furthermore the static mass and weight distribution is also entered here, together with the height of the centre of mass of the vehicle and its wheel-base length, which are required to calculate the dynamic weight distribution.

6.3.3 CVT Properties Tab

The CVT properties tab is entirely specific to the CP-CVT design, primarily allowing the dimensions of the CVT to be entered. This also serves as quick was of monitoring the physical

layout of the key components, since a plot is produced that shows the location and movement of each component as a function of α, which is controlled via a sliding bar, as shown in Figure 6-14. In addition to a plot of the key components, this tab can also display the variation of normal forces/peak Hertzian pressure, and the ball spacing, which enables the user to see the maximum number of ball elements that can be fitted into a single cage, as shown in Figure 6-15. Finally this tab also allows the user to enter the final/differential gear ratio and adjust the loading system based on a spring constant (k), and a pre-compression force (F_{A0}).

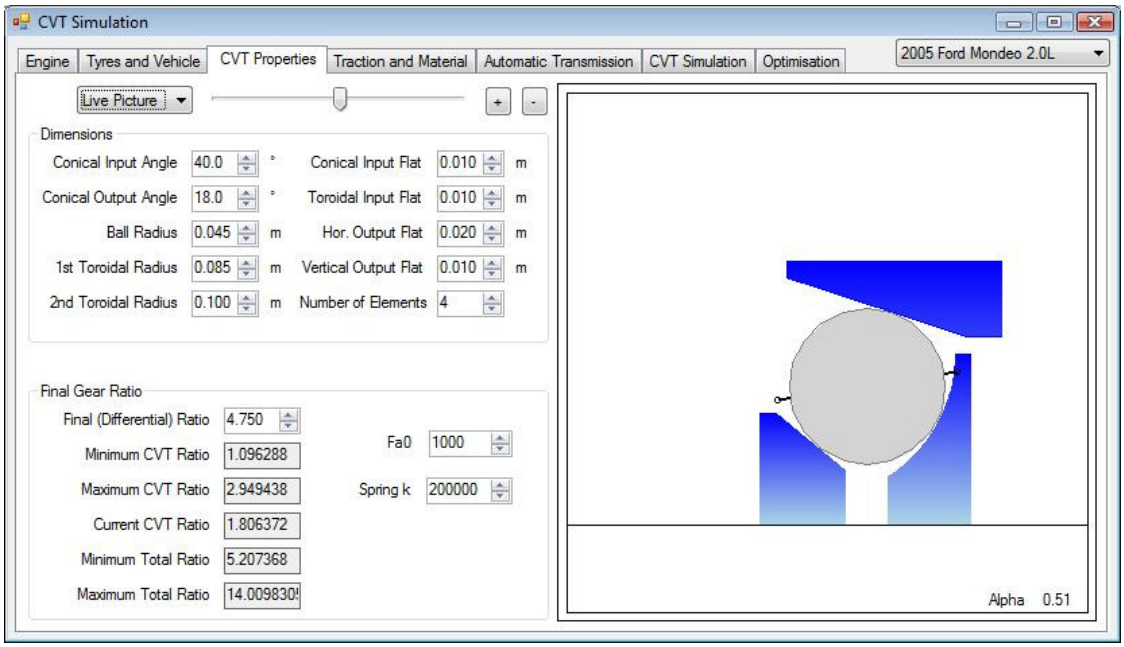


Figure 6-14: Screen-shot of CVT properties tab of simulation program

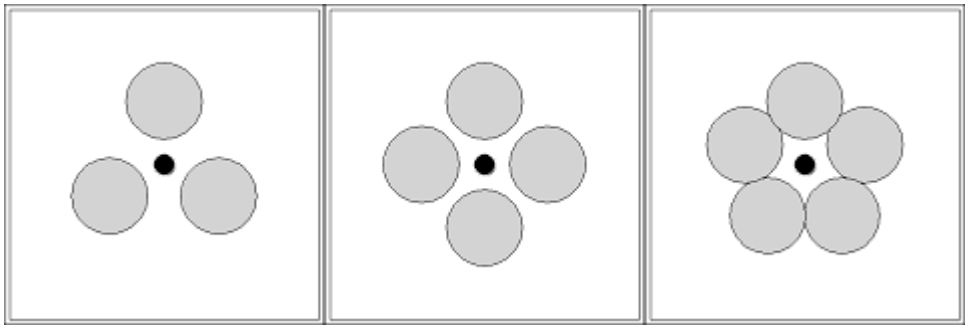


Figure 6-15: Ball spacing plots produced from simulation program

Earlier versions of the program also included the option of using a planetary gear-system, allowing 'geared-neutral' and reverse. The program automatically determined the size of gears required to achieve this geared-neutral at a particular CVT ratio, which in turn adjusted the

maximum theoretical vehicle speed in both forward and reverse directions based on the maximum engine speed, as shown in Figure 6-16.

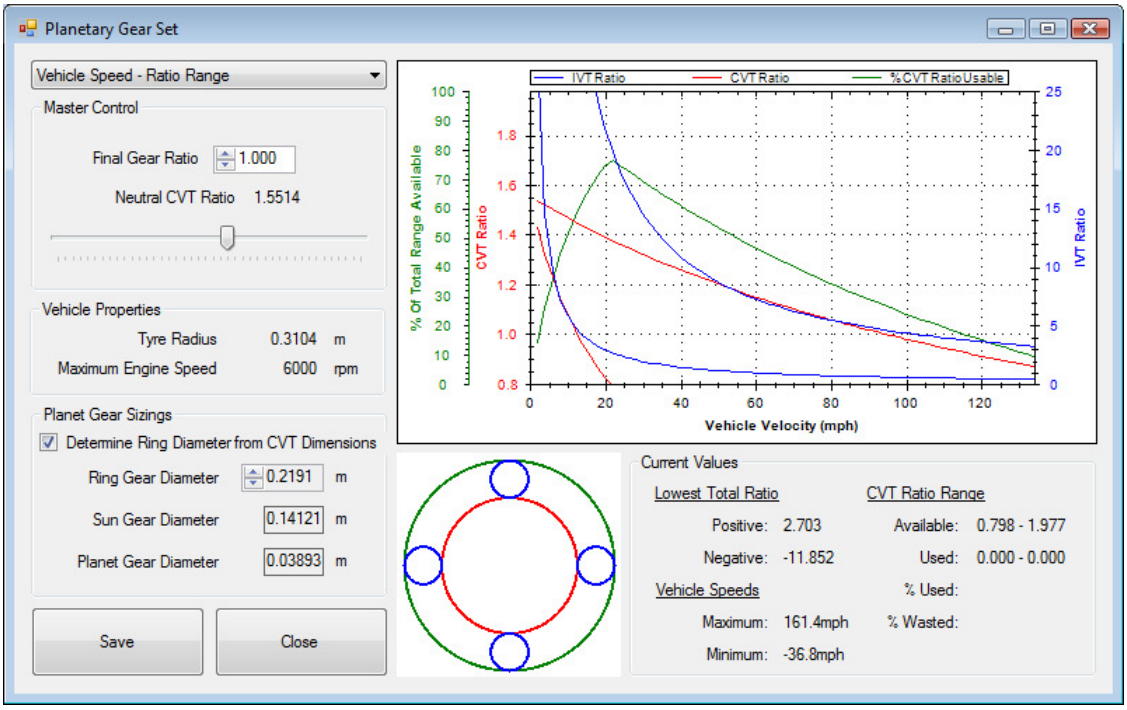


Figure 6-16: Abandoned planetary-gear system feature

Even though this aspect of the program was completed and fully functional, the difficulty of implementing a planetary gear system within the simulation with a fixed loading system required this feature to be abandoned.

6.3.4 Traction and CVT Material Tab

The traction and material properties tab allows the user to input details regarding the behaviour of the traction fluid used. Since spin effects are ignored, the main purpose of this is to determine the relationship between slide-roll ratio and traction coefficient. Whilst each parameter (such as maximum traction coefficient) can be adjusted manually, a number of predefined curves are also included based on the experimental data described earlier for a particular traction fluid (Santrotrac50) at either 40°C, 80°C, or 120°C, as shown in Figure 6-17.

Additionally this tab also contains the CVT material property parameters, such as density, elastic modulus and Poisson's ratio, which affect the overall mass, Hertzian pressure, and moment of inertia of the CVT components.

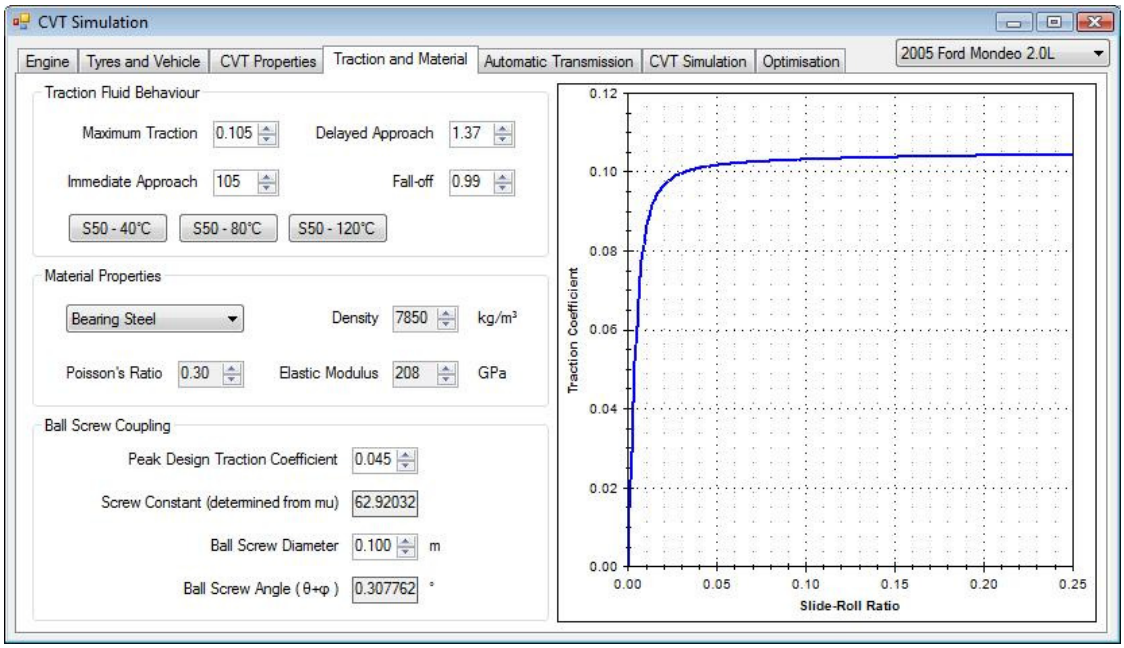


Figure 6-17: Screen-shot of traction properties tab of simulation program

6.4 Results

6.4.1 Validation of Model – Automatic Transmission Simulation

In order to verify certain aspects of the vehicle simulation, initially the CP-CVT was replaced by a simple, automated, discrete-ratio transmission. In additional to allowing a comparison with real performance values, this simulation also provides a reference for comparing the behaviour and benefits of the CP-CVT.

The transmission system is based around a simple shift-map that can be controlled manually or automatically assigned based on throttle position and wheel speed, as shown in Figure 6-18.

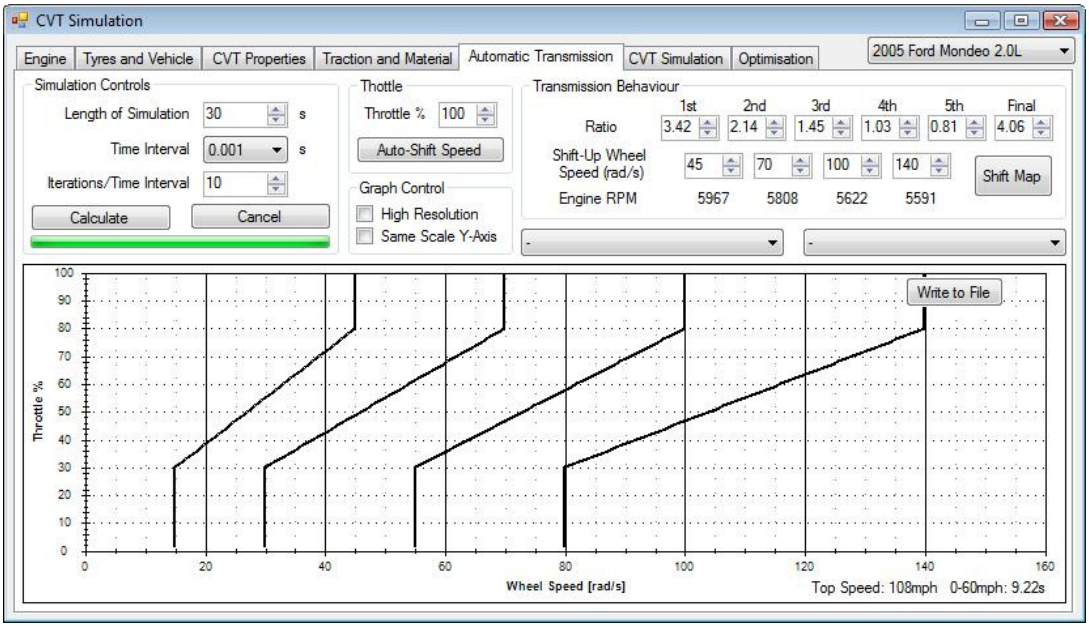


Figure 6-18: Automatic transmission shift map from simulation program

Using this simpler model, two different vehicles were simulated for which performance figures were readily available: a 2.0litre family saloon (2005 Ford Mondeo) and a 1.3litre hatchback (2001 Ford Fiesta). For each simulation, parameters such as vehicle mass, transmission ratios, and tire size were adjusted to reflect the real vehicle properties. Aspects of the simulation, such as simulation length and iteration time, are all controllable from within the program, which is also capable of displaying a number of different parameters with respect to time, including:

- Vehicle speed (mph and m/s) and acceleration (m/s^2)
- Drag force (N)
- Engine speed (rpm)
- Tire speeds, slip, torque applied and friction coefficient
- Transmission ratio

An example of the results displayed in the program of the 2005 Ford Mondeo simulation is shown in Figure 6-19, which assumes simple wide-open throttle acceleration throughout.

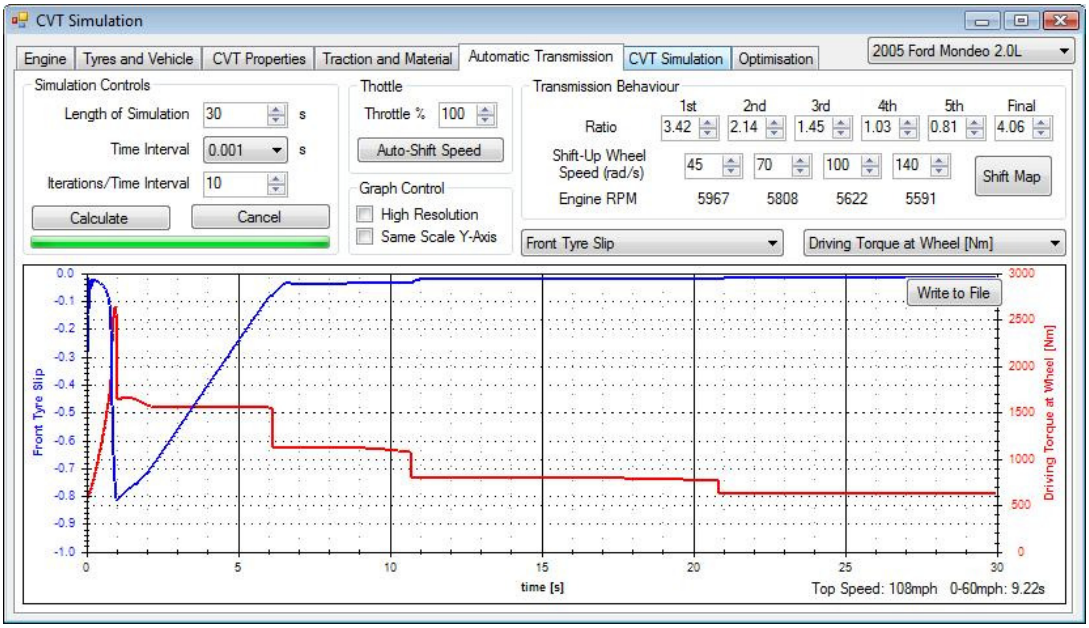


Figure 6-19: Driving force and tire slip graphs presented by simulation program

From this graph it is clear that initially a large amount of tire slip occurs, since the throttle and hence engine torque output is set to maximum (Wide-open Throttle), spinning the wheels. Gradually, as the driving torque reduces from over-revving, the tire's tangential speed begins to match that of the road surface and hence slip decreases, as expected.

In addition to displaying these properties within the program, it is also possible to write the data to a Microsoft Excel file, as shown in Figure 6-20, which shows the vehicle speed in m/s and transmission ratio change of the same simulation over 30 seconds.

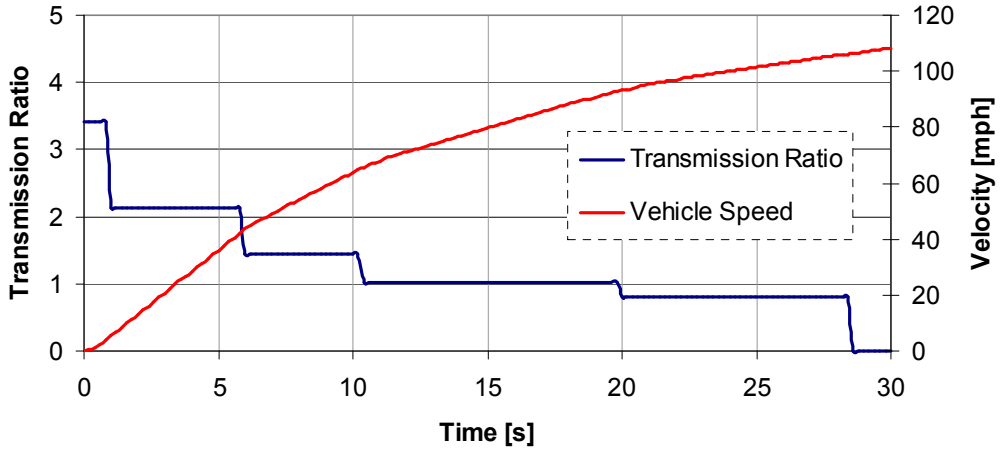


Figure 6-20: Simulated acceleration of step-gear vehicle

According to these figures, the simulation performed as expected, which was further verified by a comparison of the simulated and real vehicle performance as shown in Table 25.

Table 25: Comparison of simulated and real data

	0-60mph [s]		*Top Speed [mph]*	
Vehicle	*Real*	*Sim.*	*Real*	*Sim.*
2.0l Family saloon (107kW @ 4,500rpm) (190Nm @ 6,000rpm) 1400kg mass	9.5	9.22	135	132.3
1.3l hatchback (43kW @ 2,500rpm) (101Nm @ 5,000rpm) 1200kg mass	18	17.08	95	95.5

The good correlation of the real and simulated data can be seen to validate the simulation. The simulated top speeds (which required a longer simulation) were almost identical to the manufacture-stated values, whilst the difference in acceleration performance could be due to the extra time taken to change gears and disengage/engage the clutch, which was not included in the simulation.

6.4.2 Results of CVT Simulation

6.4.2.1 Comparison with Automatic Transmission Simulation

The targets set in the genetic algorithm optimisation (Chapter 5) coincide with the engine data for the 2005 Ford Mondeo, and hence the CP-CVT simulation was run using the component dimensions determined previously for a typical family car (β=19.7°, γ=56.8°, R=0.0610m, R_1=0.0686m, r_0=0.0891m)

The loading system (spring-constant and pre-compression) was initially adjusted to enforce a near-constant torque input of 170-180Nm. It was found however that even with an iteration time of 0.001s, there was a large amount of instability in the movement of the toroidal disc. As shown in Equation 6.24, the movement of the toroidal disc is calculated based on the immediate acceleration and velocity, However it was found that these values are too large and hence the movement fluctuates between different operating points as shown in Figure 6-21.

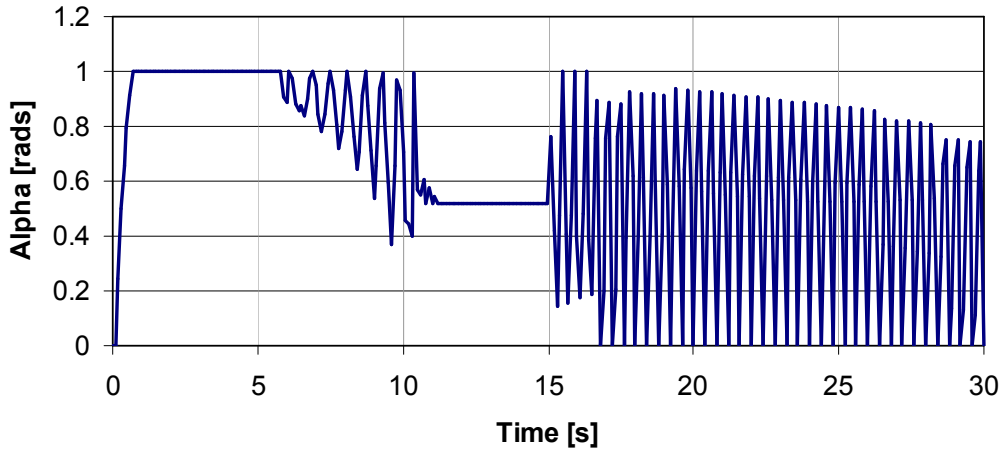


Figure 6-21: Instability of toroidal disc position

This problem was rectified by artificially increasing the mass of the toroidal disc for the purposes of linear acceleration calculations, thus decreasing the magnitude of the acceleration. Using this modification, a comparison was made between the performance of the CVT transmission and the automatic transmission described previously (Figure 6-22), again in wide-open throttle conditions (WOT).

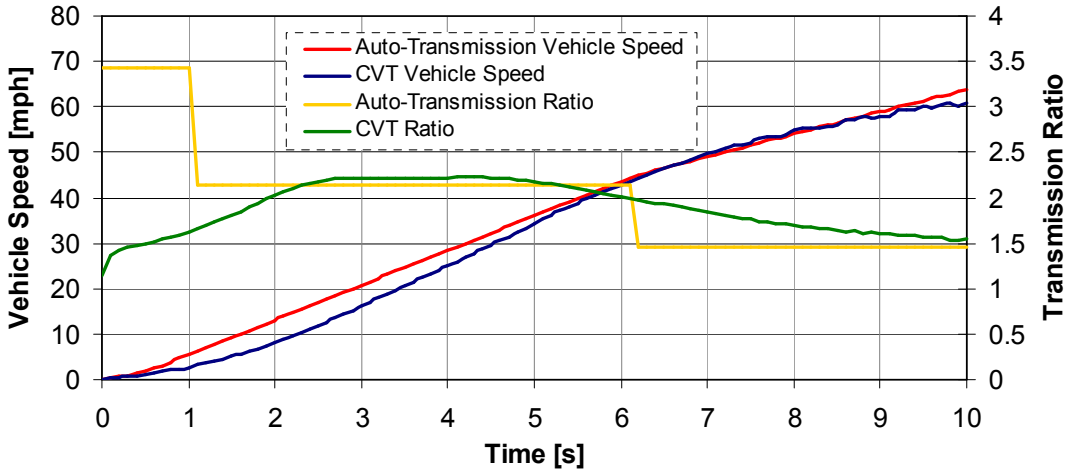


Figure 6-22: Comparison of CVT and automatic transmission performance

Compared to the discrete-ratio simulation run previously, the CP-CVT offered similar performance when accelerating at full throttle, with the added benefit of a more stable engine speed and smoother acceleration. Previous research found an improvement in acceleration when using a CVT (Boos and Mozer, 1997), however this is not reflected here. This is possibly due to inefficient gear ratio usage resulting from a fully automated torque-controlled transmission ratio.

6.4.2.2 Influence of Loading Spring Properties

From Figure 6-22 it can be seen that initially the ratio of the CP-CVT is at its lowest value (α=0). As the torque from the engine rises the ratio is automatically adjusted until it reaches its upper limit, which takes approximately 2.5seconds. The transmission ratio then falls slowly in response to the reduced acceleration of the vehicle, and hence reduced torque applied to the ball screw coupling. Although this time is comparable to other traction CVTs (Fuchs, Hasuda, and James, 2000), the initial response time can be reduced through optimisation of the loading system in order to improve the initial acceleration:

- Figure 6-23 shows the effect of increasing the initial compression (F_{A0}) of the spring, which leads to a longer response time (approximately 5 seconds), and therefore the initial acceleration is reduced, but at higher speeds the full range of ratios is used and hence the final velocity is larger (87mph).

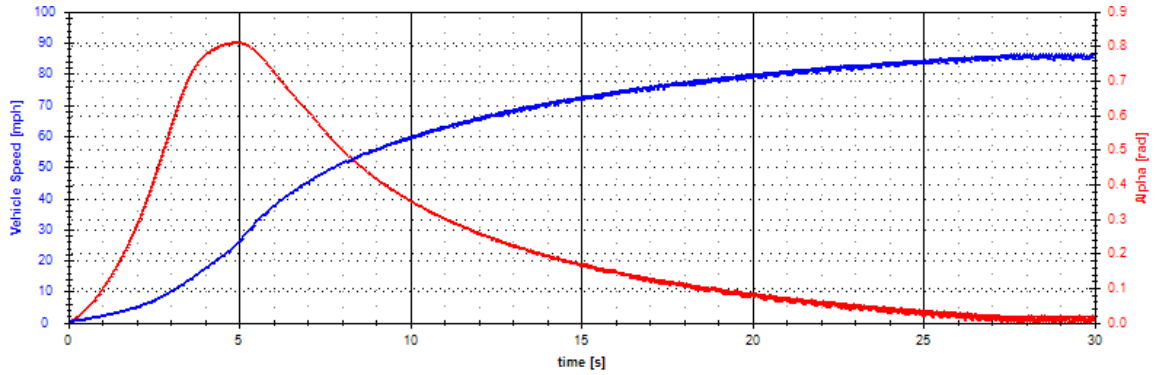


Figure 6-23: Alpha response and vehicle speed with increased spring pre-compression

- Figure 6-24 shows the effect of decreasing the initial compression of the spring, which leads to a shorter response time (approximately 2 seconds), and therefore the initial acceleration is increased, but at as the vehicle velocity increase the transmission ratio does not reach its lowest value, and hence the final velocity is lower (70mph).

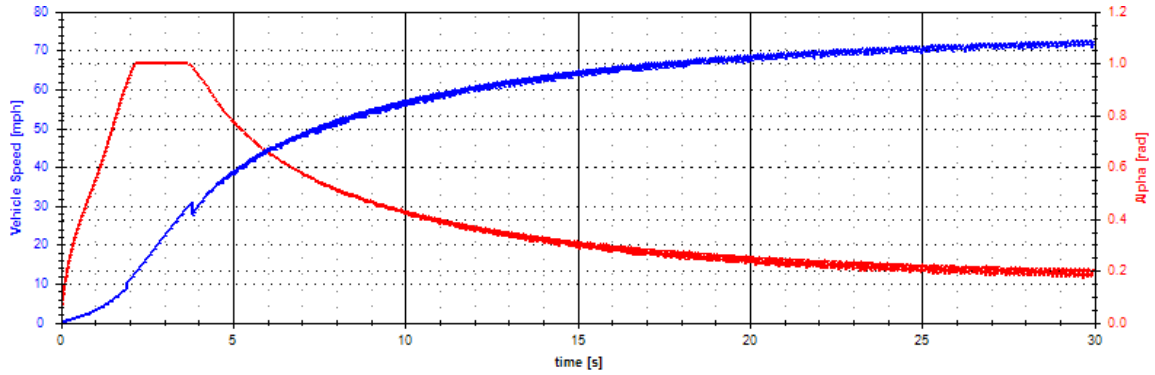


Figure 6-24: Alpha response and vehicle speed with reduced spring pre-compression

- Figure 6-25 shows the effect of decreasing the initial compression of the spring, but increasing its stiffness (k), which means the CVT still responds relatively quickly (2.5 seconds), but at higher velocities a wider range of ratios is employed, and hence the final velocity is also higher (75mph).

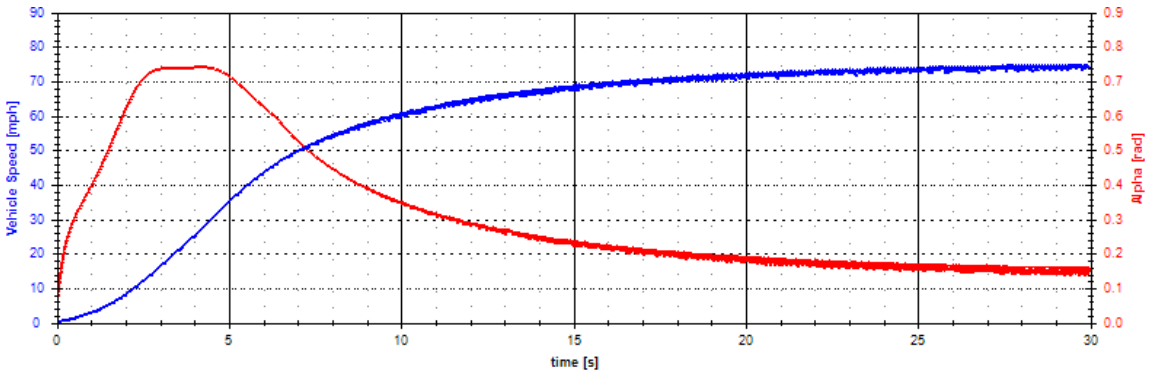


Figure 6-25: Effect of reduced spring pre-compression but increased stiffness

6.4.2.3 Influence of Power-Variation Coefficient

It was assumed during the optimisation process that the best ratio response from the CVT can be achieved by having a linear relationship between the force required at the toroidal disc and the movement of the toroidal disc (Section 5.1.2.7). In order to determine if this assumption was correct, the optimisation process described earlier was repeated with identical target except for the power variation coefficient for which the target was set to 0 rather than 1 (indicating a higher variation in power output). The results produced by the algorithm are shown in Table 26.

Table 26: Comparison of low and high power variation correlation optimised properties

Technical Characteristic	*High P.V Correlation*	*Low P.V Correlation*
Overall Efficiency	91.9%	92.43%
Normalised Transmission Ratio	4.05	4.91
Maximum Torque In	180.4Nm	216.0Nm
Indicative Mass	31.1kg	24.7kg
Length	0.13	0.13
Diameter	0.14	0.14
Power Variation Coefficient	0.97	0.00

It is difficult to compare the two sets of parameters directly, since each set of dimensions that yielded these characteristic would require different loading characteristics to compensate for the change in geometry and different final gear ratios to ensure a similar maximum transmission

ratio is available. It was generally found that the CP-CVT with the higher power variation coefficient (low variation in power output) performed better when accelerating from rest as it was able to use a wider range of ratios. However it was also found that the CP-CVT with the lower power variation coefficient was dynamically more stable (at least within the simulation). The reason for this is that the higher power variation coefficient implies that the CVT has a wider range of torque output for a fixed torque input. Hence for any particular engine output there are a vast number of different ratios the CP-CVT could operate at (this is after all, the rationale of constant power). The dynamic instability could hence be seen as a fault with the discrete numerical simulation rather than the design. The proof of the power-variation concept is shown in Figure 6-26, which shows the CVT transmission ratio compared to the transmission ratio required to maintain a constant engine speed (ideal ratio). This correlation could only be achieved with the dimensions that demonstrated a higher power variation coefficient.

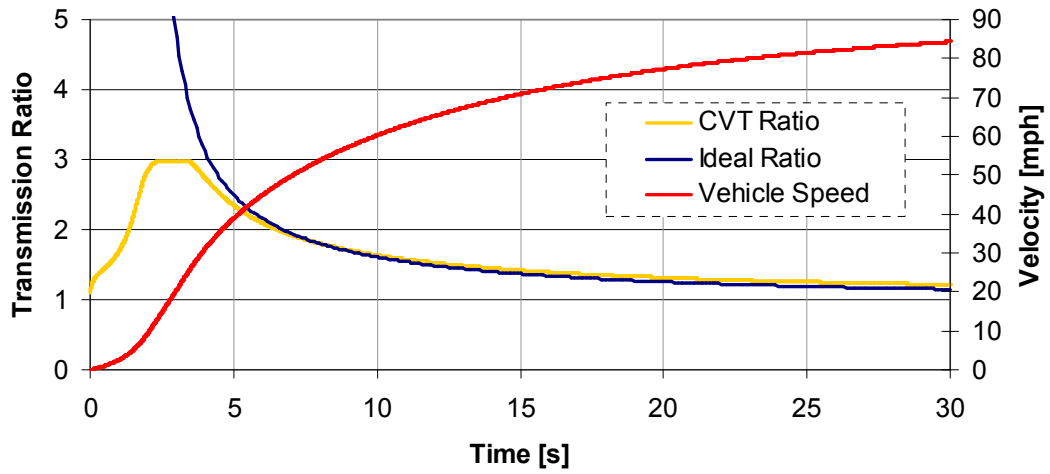


Figure 6-26: Simulated acceleration of CVT vehicle

Ratio control has formed a large part of the research into continuously variable transmissions recently. To fully take advantage of the non-discrete ratios a CVT offers the full behaviour of the engine must be understood. The transmission ratio can then be adjusted to force the engine to operate at the required speed and torque output as often as possible. It is generally assumed, and has been proven (Pfiffner and Guzzella, 2001) that the highest fuel economy can be achieved initially through a constant engine speed and higher torque output, hence the ideal ratio shown in Figure 6-26 simply assumes a constant engine speed. From this it is clear that initially the CVT cannot provide the required ratio without the use of an epicyclical gear-set, however once the engine speed reaches the desired speed the CVT automatically provides the almost precisely the correct ratio. It is believed that the minor variation from the ideal ratio curve is a result of the marginal imperfection in the power variation coefficient (0.97).

6.4.2.4 Decelerating and Cruising

The simulations thus far have only looked at the behaviour of the CP-CVT in a vehicle when accelerating from rest. We will now look at a situation where a vehicle initially accelerates from rest with only partial throttle, then reduces engine power for a short time, and then begins to accelerate again, as shown in Figure 6-27. The throttle is controlled automatically based on the vehicle and engine speed at each moment, hence emulating a driver's response.

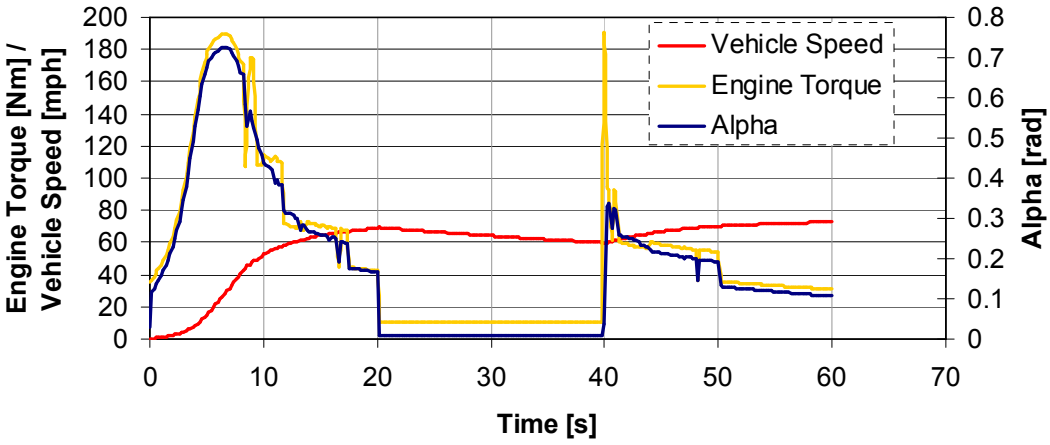


Figure 6-27: Dynamic driving situation simulation

The CVT now takes longer to reach the higher transmission ratio, which occurs because of the reduced throttle conditions. By initially only using partial throttle, the CP-CVT uses a lower transmission ratio (equivalent to a higher gear in a manual transmission), which lowers the engine speed and potentially increases fuel efficiency. After 20seconds, the throttle is released fully, reducing engine torque output immediately, and hence the vehicle velocity begins to reduce as drag and rolling resistance forces exceed the driving force. After 40 seconds, the throttle is again applied, but this time it is done sharply, which results in an instant increase in engine torque and response from the CP-CVT, providing an immediate increase in acceleration and hence velocity.

To ensure the engine does not rotate too quickly, the transmission ratio available at this increased velocity is limited. To ensure this doesn't happen, the throttle is simply reduced once the engine speed exceeds a predefined value (6,000rpm in this situation).

Throughout the entire driving scenario, the correlation of engine torque and α implies that the transmission ratio is in reality controlled by the engine torque, rather than the driving scenario.

6.5 Conclusions

The vehicular simulation created in this chapter has been shown to have a good correlation with real vehicle behaviour when coupled with a standard discrete ratio transmission. By using manufacturer-obtained values such as engine torque, wheel sizing, and gear ratios, a very close approximation was obtained in both the theoretical acceleration and top speed of the vehicle. By utilising the optimised dimensions of the CP-CVT, a comparable vehicular performance was found in terms of acceleration and top speed. Whilst certain literature sources claim that a 10% increase in acceleration can be achieved through the use of a CVT (Boos and Mozer, 1997), this requires complete control of the CVT's transmission ratio to ensure that the maximum torque available from the engine is used as often as possible, which is not easily achievable with an entirely torque-controlled CVT. Despite this the CP-CVT performed admirably, providing a suitable gear ratio that reflected engine torque, throttle position and vehicle loads. Generally it was found that the CP-CVT's transmission ratio and hence engine's operating curve was dictated by engine torque and throttle position, either providing maximum torque output and acceleration or lower engine speeds and thus increased efficiency.

A large part of the behaviour of the CP-CVT depends on the force applied to the toroidal disc. When using a fixed spring-type loading system, Figure 6-23 to Figure 6-25 show that selecting an appropriate spring stiffness and pre-compression is perhaps even more critical than the component dimensions to the operation and response of the CP-CVT. Generally it was found that a very low or even zero pre-compression provided the best results since it allows the CVT to reach the highest gear ratio quicker, whilst a stiffer spring ensures that a wider range of transmission ratios are utilised. However despite this, the choice of spring stiffness depends largely on the desired cruising speed, since it affects how quickly the CP-CVT reaches its maximum transmission ratio. If it reaches this ratio too quickly then acceleration at higher speeds suffers, whilst if it doesn't reach this ratio at all then the overall acceleration of the vehicle is reduced through inefficient ratio usage. Previous research by Cretu and Glovnea (2005) discussed the use of a hydraulic or similar loading system, which would allow a greater control of the ratio, improving both acceleration performance and engine efficiency. This would however significantly complicate the design and operation of the CP-CVT, which was found to perform well, even when fully autonomous.

CHAPTER 7: COMPLEMENTARY RESEARCH: RUBBER FRICTION

Parts of this chapter will be presented at the 2010 International Tribology Congress - ASIATRIB 2010 and published in the associated proceedings.

7.1 Introduction

7.1.1 Chapter Summary

Predicting and modelling the behaviour of the tire-road contact is seen as crucial in increasing the performance of vehicular breaking systems. Determining the optimum breaking conditions has become even more important as nearly all vehicles manufactured today are fitted with Anti-lock Braking Systems (ABS), which automatically adjust the braking force applied to the wheel to ensure that the slip is minimised and the friction force is as high as possible. Reducing slip is crucial to increasing vehicle control and reducing tire wear. A better understanding of tire friction behaviour can help to improve the effectiveness of these types of safety systems. Tire friction models typically incorporate a friction coefficient (μ), which can be defined as the ratio of friction force to normal load. By monitoring the factors that influence μ, improvements can be made to vehicular safety when braking, and vehicular performance when accelerating, especially in lower friction conditions such as rain or snow.

Typically fundamental friction models that incorporate finite element methods (FEM) have not yielded any significant findings that could be incorporated into tire friction models; hence current models tend to be simple curves, which are fitted to experimentally obtained results.

This chapter attempts to use controlled laboratory frictional experiments in order to determine the physical validity of existing models, with a focus on the effect of macroscopic roughness and whether particular consistent roughness can be incorporated into existing friction models.

7.2 Existing Tire Friction Models

7.2.1 Pacejka's Friction Model

The parameter that defines the magnitude of friction in most tire friction models is the slip ratio, which is simply defined as the difference in velocities between the road surface and the surface of the tire. Although there are several different mathematical definitions of slip, it is generally defined as shown in Equation 7.1.

$$s = \frac{|v - \omega r|}{\max(v, \omega r)} \tag{7.1}$$

Where s = slip ratio, v = vehicle velocity, ωr = tangential speed of the tire.

This definition of slip assumes that the magnitude of the friction is independent of whether the vehicle is accelerating or breaking, and is instead only dependent on the *magnitude* of the slip between the tire and the road surface. One of the most widely accepted models of friction coefficient as a function of slip is the Pacejka's friction model, which states:

$$\mu = D\sin(C\arctan(Bs - E(Bs - \arctan(Bs)))) \tag{7.2}$$

Where B, C, D and E are constants, which are chosen to match experimental data.

Whilst attempts have been made to quantify or apply physical meaning to these constants (Bakker, Nyborg and Pacejka, 1987), they are generally considered to be vague. The formula's popularity is hence its simplicity and ability to match experimental data. One of the largest criticisms of this model is that the constants must be changed when the vehicle load or velocity change, both of which have been shown to significantly affect the magnitude of the friction coefficient. Typically, the effects of both velocity and load are relatively predictable, affecting the maximum point of the friction curve, rather than the shape, hence if the friction curve is known for a particular load and velocity, it can be shifted with reasonable accuracy for other loads and velocities.

A typical curve produced by this model is shown in Figure 7-1. This graph shows experimentally obtained results for a friction test rig on a smooth surface at two different speeds (▲= 20mm/s; X= 10mm/s) together with fitted curves produced using the 'magic formula'. One of the most important characteristics of this curve is the peak value of friction that occurs at relatively low slip, together with the falloff that occurs at higher slip values. This is the area of the curve that is of most interest to ABS and TCS designers as operating at a slip ratio that allows a higher value of the friction coefficient can significantly reduce the stopping distance or improve the control and stability of a vehicle. Typically the reduction in friction that occurs at

higher slip values is intensified at higher speeds, normally due to the increased heat generated, which affects the tire material.

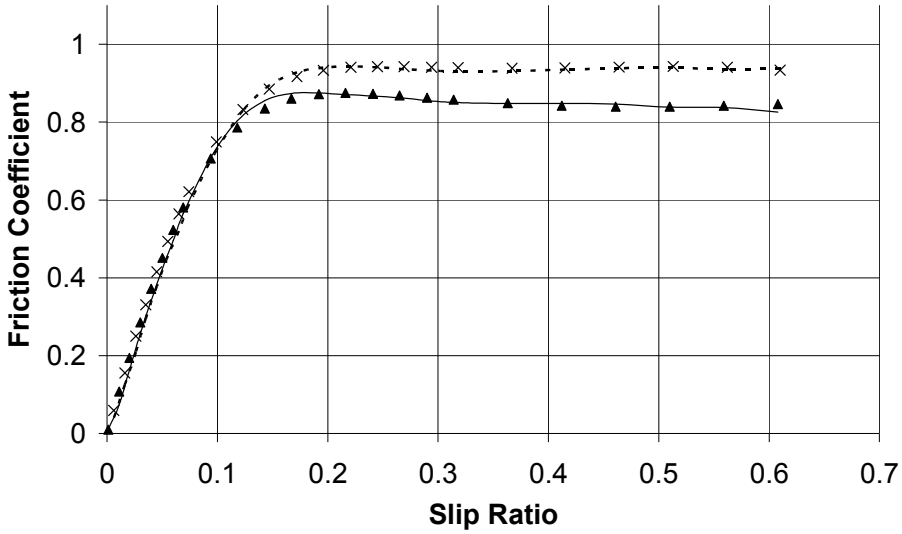


Figure 7-1: Typical 'magic formula' curves fitted to experimentally obtained curves

Whilst Pacejka's formula offers a convenient model of tire friction, it offers very little explanation of the physical intricacies that occur in the tire-road contact.

7.2.2 Dahl Friction Model

The Dahl friction model (Dahl, 1976) is based on the concept of displacement. This model assumes that the magnitude of the friction force is a function of tire surface displacement, similar to a stress-strain curve, as shown in Figure 7-2.

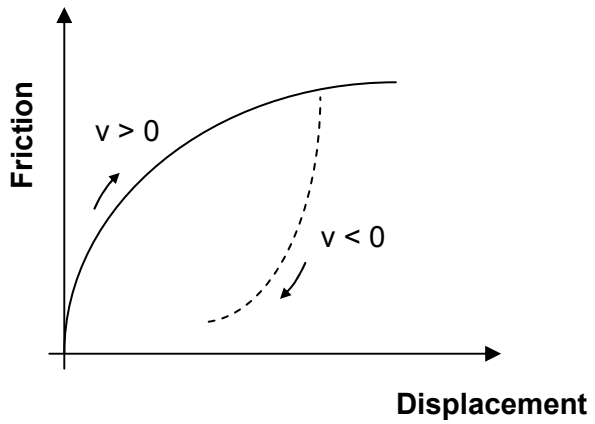


Figure 7-2: Stress-strain relationship that forms basis of Dahl friction model

Whilst this model is making an assumption that friction responds in a similar way to a simple deflecting beam, it is based on a physical phenomena, rather than experimental graph fitting as in Pacejka's friction model. In its simplest form this model takes the following form:

$$\frac{dF}{dx_r} = \sigma_0 \left(1 - \frac{F}{F_c} \right) \tag{7.3}$$

Where F = the friction force, F_c = maximum friction force, σ_0 = stiffness coefficient (the slope of the curve at $x_r=0$), and x_r = the relative displacement. This model can be easily adapted to include relative velocity since it can be seen that $dx_r/dt = v_r$, hence:

$$\frac{dF}{dt} = \frac{dF}{dx_r}\frac{dx_r}{dt} = \frac{dF}{dx_r} v_r = \sigma_0 \left(1 - \frac{F}{F_c} \right) v_r \tag{7.4}$$

7.2.3 LuGre Friction Model

An extension of Dahl's model that better fits observed phenomenon has been proposed by LuGre (Deur, 2001). In this model, friction is modelled as the average deflection force of elastic bristles. The model uses the fact that if a tangential force is applied to bristles, they will deflect in a predicable manner, similar to a spring. If the force, and therefore the deflection, is large enough the bristles will begin to slip. In steady state conditions, the average deflection will be a function of the slip velocity (Olsson, et al. 1998). The simplest form of this model in terms of a friction force (F) is shown in Equation 7.5, which assumes a point load distribution.

$$F(v_r) = \left(\sigma_0 z + \sigma_1 \dot{z} + \sigma_2 v_r \right) F_n \tag{7.5}$$

Where

$$\dot{z} = v_r - \frac{\sigma_0 |v_r|}{g(v_r)} \tag{7.6}$$

And:

$$g(v_r) = \mu_c + (\mu_s - \mu_c) e^{-|v_r / v_s|^{\alpha}}$$

Where:

- σ_0 is the rubber stiffness
- σ_1 is the rubber lumped damping
- σ_2 is the viscous relative damping
- μ_c is the normalised Coulomb friction
- μ_s is the normalised static friction
- v_s is the Stribeck relative velocity.

Given that the relative velocity (v_r) can be defined as $v_r = r\omega - v$, and that $\mu = F/F_n$ this model can easily be modified to show friction as a function of slip thus making it comparable to Pacejka's model.

This model assumes a point load, which is not particularly accurate. If the contact patch of length L is divided into a number of infinitely small elements ($d\zeta$), and the model is assumed to be steady state with respect to time, then the total friction can be calculated as shown in Equation 7.7.

$$F = \int_0^L dF(\zeta) d\zeta \tag{7.7}$$

Whilst the friction produced by each individual element (bristle) can be calculated from Equation 7.8.

$$dF(\zeta) = (\sigma_0 z(\zeta) + \sigma_2 v_r) dF_n(\zeta) \tag{7.8}$$

Where the steady-state value of z can be calculated from Equation 7.9.

$$z(\zeta) = \frac{g(v_r)}{\sigma_0}\left(1 - \exp(-\frac{\sigma_0}{g(v_r)}\left|\frac{v_r}{\omega r}\right|\zeta\right) \tag{7.9}$$

Combining Equations 7.7-7.9 yields an expression for the friction force as a function of relative velocity (Equation 7.10)

$$F = \int_0^L \left[g(v_r)\left(1 - \exp\left(-\frac{\sigma_0}{g(v_r)}\left|\frac{v_r}{\omega r}\right|\zeta\right)\right) + \sigma_2 v_r \right] f_n(\zeta) d\zeta \tag{7.10}$$

Where $f_n(\zeta)$ is the normal force per unit length as a function of ζ. This can be any function that can be integrated along the patch length (L) indicating the normal load distribution. It should also be noted that in order to calculate the friction coefficient, the total normal force (F_n) must be known, which can be calculated as follows:

$$F_n = \int_0^L f_n(\zeta) d\zeta \tag{7.11}$$

The simplest normal load distribution assumes a continuous load across the contact patch, i.e:

$$f_n(\zeta) = F_n / L$$

Combining with Equation 7.10 and rearranging in terms of μ:

$$\mu = \int_0^L \left[g(v_r)\left(1 - \exp\left(-\frac{\sigma_0}{g(v_r)}\left|\frac{v_r}{\omega r}\right|\zeta\right)\right) + \sigma_2 v_r \right] L^{-1} d\zeta \tag{7.12}$$

Hence:

$$\mu = g(v_r)\left(1 - \left(\left|\frac{\omega r}{v_r}\right|\frac{g(v_r)}{L}\right)\left(1 - \exp\left(-\left|\frac{v_r}{\omega r}\right|\frac{L\sigma_0}{g(v_r)}\right)\right)\right) + \sigma_2 v_r \tag{7.13}$$

A comparison can now be made between the LuGre and Pacejka models, as shown in Figure 7-3, which uses the constants shown in Table 27.

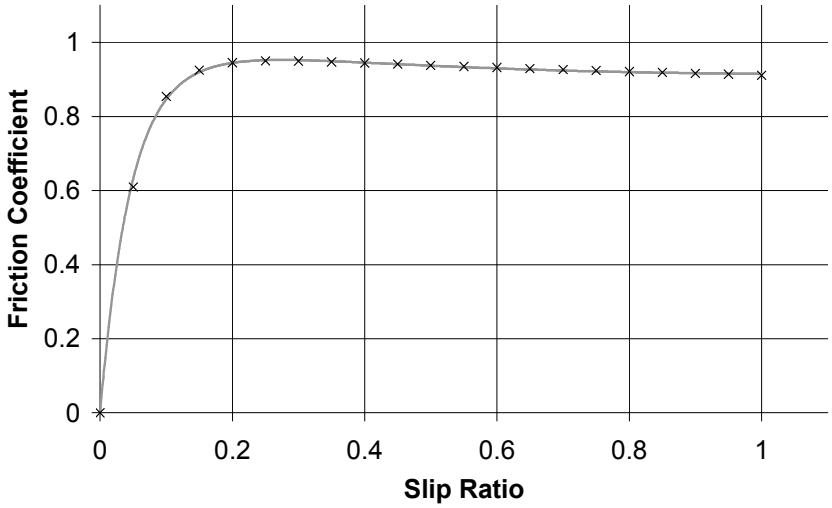


Figure 7-3: Comparison of LuGre (solid line) and Pacejka (crosses) friction models

Table 27: Pacejka and LuGre constants used for curves shown in Figure 7-3

Pacejka Constants		*LuGre Constants*	
B	7.6	L	0.2 m
C	1.8	σ_0	180 m^{-1}
D	0.95	σ_2	0.002 s/m
E	0.97	μ_c	0.8
		μ_s	1.55
		v_s	6.5 m/s

Figure 7-3 shows that the LuGre model is fully capable of replicating curves obtained through the Magic formula, whilst still maintaining a foundation in physical, measurable parameters.

7.2.4 Limitations of the LuGre Model

Assuming a uniformly distributed load, the effect of a change in patch length is shown in Figure 7-4.

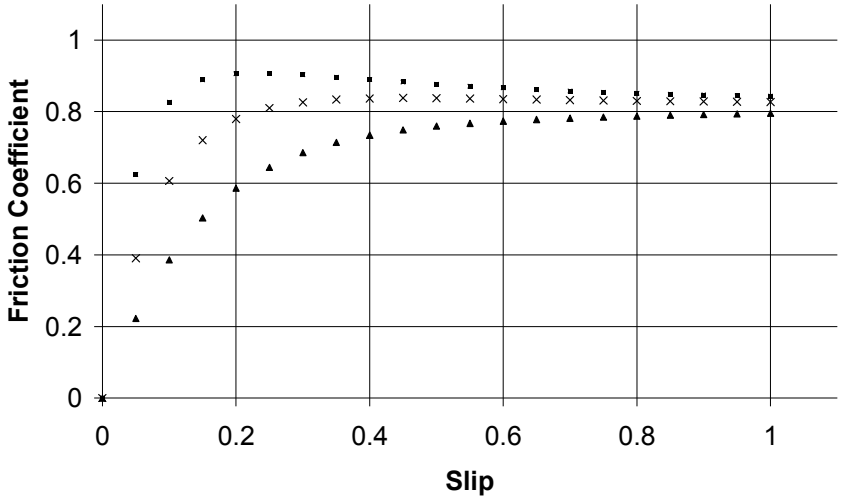


Figure 7-4: Effect of patch length on LuGre friction (■ = 0.2m; x = 0.1m; ▲= 0.05m)

From this figure it can be seen that according to the LuGre model, a reduction in patch length dramatically changes the overall shape of the curve for a fixed stiffness coefficient. This contradicts Amonton's fundamental law of friction and several results (Krim, 1996), which state that the apparent contact area, and hence patch length have no affect on the friction coefficient.

Despite this apparent limitation of the model in explaining physical phenomenon, it still remains a widely accepted and reasonable starting place in attempting to explain the effect of measurable changes on friction coefficient.

7.3 Experimental Methodology

7.3.1 Tire Friction Test Rig

In order to determine the influence of the parameters in the LuGre model and see how these compare to the LuGre model, several road surface samples were produced each with different macroscopically rough profiles, thus altering the normal load distribution and contact pressure. These samples were tested on a specially designed tire-friction test rig, shown in Figure 7-5. This scaled rig, (which was designed and built by researchers at Kyushu University in Japan, Morita et al., 2010), uses a 64mm diameter solid rubber tire and allows a variety of different velocities, loads and slip values to be tested, all of which can be independently controlled. The rubber tire is driven by a stepper motor capable of accurately measuring and controlling the rotational speed. The normal load is controlled by altering the weight applied to the tire, which is attached through a linear bearing that allows vertical movement. The vertical deflection is measured through a proximity sensor. Longitudinal motion and velocity are controlled through a ball screw and motor, whilst longitudinal forces are measured via strain gauges attached to the vertical arms holding the tire, as shown in the Figure 7-5.

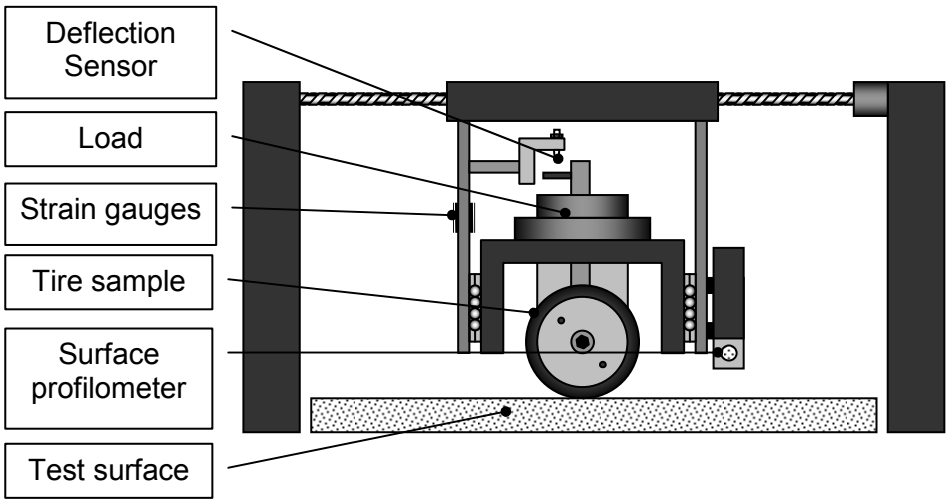


Figure 7-5: Roughness and friction measurement test rig

7.3.2 Test Surfaces

Tests were conducted on various microscopic and macroscopic roughness profiles. Microscopically rough surfaces consisted of various abrasive papers of known particle sizes. In order to determine the effect of 2-dimensional macroscopic asperities (uniform surface profiles across the width of the contact), a number of moulded cement samples were produced incorporating various combinations of asperity radius and pitch. This allowed a controllable

number of contact points and profiles whilst maintaining a consistent microscopic roughness. Four main samples were used, using either a 1mm or 2mm asperity radius and 2mm or 4mm pitch. This combination allowed the effect of both pitch and radius to be studied independently as shown in Figure 7-6.

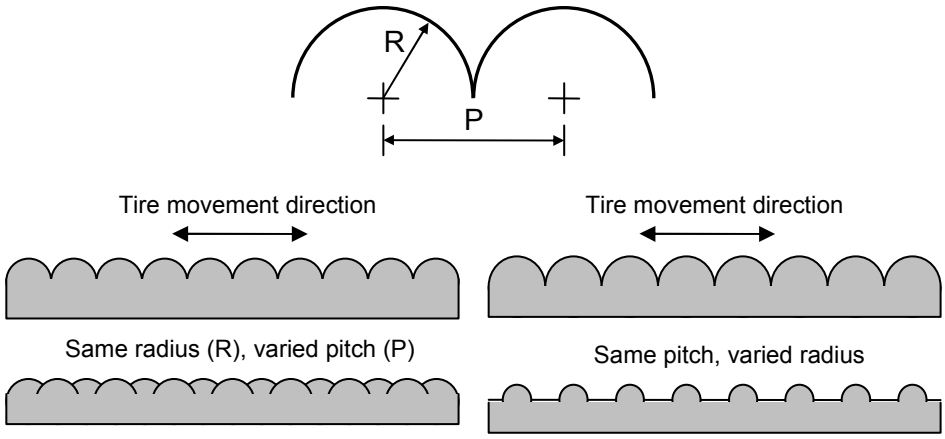


Figure 7-6: Examples of macroscopic roughness profiles

7.3.3 Test Conditions

Tests were conducted multiple times on each sample in order to ensure reliable results and exclude any erroneous results. Consistency was ensured by cleaning samples before each test, and removing any rubber particles that had been left on the cement surface. The surface of the rubber tire was also smoothed and cleaned to ensure that the surface remained approximately constant for each test.

Each test consisted of measuring the deflection of the supporting beam which was calibrated against a known longitudinal force. By knowing the normal load applied in each test, the friction coefficient could be calculated. Friction was measured along 300mm of surface in each direction for every slip value, with the average value of friction being taken. All tests were conducted at 10 mm/s and 20mm/s.

Although initially each test was carried out with several different normal loads, it was found that there was very little measurable variation in friction coefficient between different loads, and in any case no particular pattern that would yield any useful findings. Fundamental theoretical research has shown that true contact area is “very nearly proportional” to the normal load

(Archer, 1957), and furthermore that the coefficient of friction decreases markedly with increased contact pressure (Denny, 1953). This is only true up to a limit however; once the normal load is sufficiently high, the real contact area no longer increases and friction force remains constant (Denny, 1953). Because of this complex behaviour, the tests were all conducted at a fixed normal load (19.6N), focusing instead on the nature of the contacting surfaces. A summary of the test conditions is shown in Table 28.

Table 28: Summary of test conditions

Tire	*Thickness*	10mm
	Radius	32mm
	Elastic Modulus	0.05 GPa
	Poisson's Ratio	0.5
Cement Surface	*Elastic Modulus*	30 GPa (typ.)
	Poisson's Ratio	0.28 (typ.)
Test Conditions	*Normal Load*	19.6N
	Speed	10mm/s or 20mm/s
	Running Length	300mm
	Slip Ratio	Approx. 0-0.7

7.4 Results

7.4.1 Microscopic Roughness

Although macroscopic roughness is of more interest in terms of the friction models, initially tests were conducted on various microscopically abrasive papers, the results of which are shown in Figure 7-7. The three papers tested were A400, A240 and A120 according to the CAMI Grit designation, which correspond to average particle sizes of 23µm, 53µm and 115µm respectively.

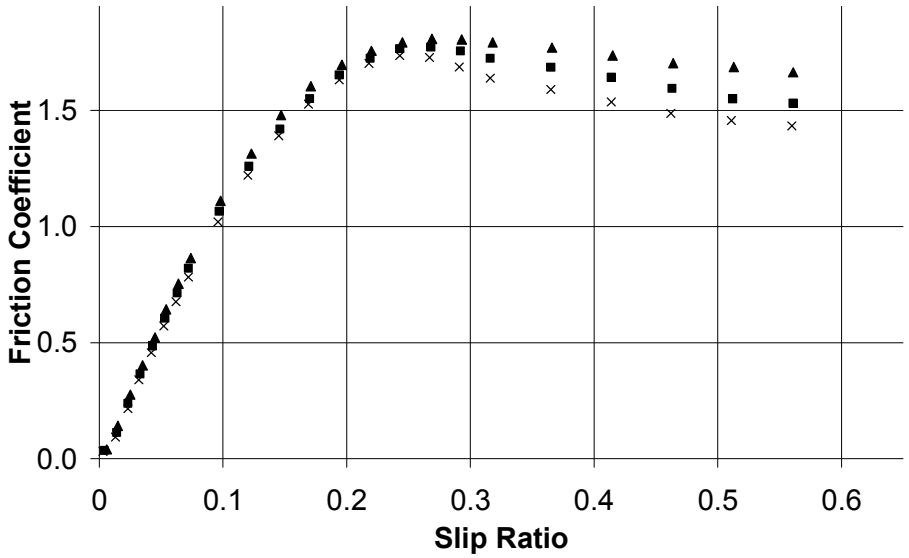


Figure 7-7: Effect of microscopic roughness at 20mm/s (▲= A120; ■ = A240; A400 = x)

As shown in this figure, increasing the particle size of microscopic roughness also increases the friction coefficient. This is perhaps expected and is analogous to using a coarser sandpaper to smooth a surface faster. This also implies that the primary friction mechanism that occurs in microscopic roughness is wear. This assumption is supported by the similarities of the friction curves at low slip ratios, where wear is less dominant. Conversely at higher slip ratios, where wear is the dominating friction mechanism, the difference in friction between particle sizes is far more obvious.

7.4.2 Macroscopic Roughness

7.4.2.1 Effect of Radius of Macroscopic Asperities

One of the advantages of using a mouldable material is that both asperity radius and pitch can be controlled independently. Figure 7-8 shows the effect of doubling the asperity radius from 1mm to 2mm, whilst keeping the pitch length constant at 2mm.

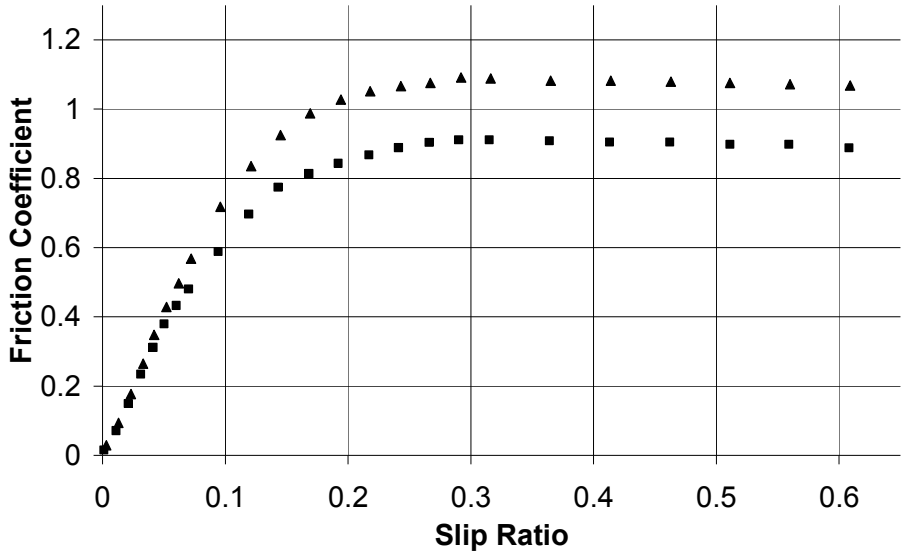


Figure 7-8: Effect of asperity radius at 10mm/s (▲=R1P2; ■ = R2P2)

At this load (19.6N), the contact is not completely conformal for either radii. These results show that as the asperity radius decreases, the average friction force increases. This is intuitively what is expected, since a smaller, sharper asperity would lead to an increased contact pressure. Although the increased friction is expected, the magnitude of the increase is perhaps lower than expected given the relative radii of the contacting surfaces. Looking at the geometrical mean radii (an important factor in determining contact pressure), of the tire and a single asperity:

Asperity radius of 1mm: $R = \frac{1}{\frac{1}{1} + \frac{1}{32}} = 0.97mm$

Asperity radius of 2mm: $R = \frac{1}{\frac{1}{2} + \frac{1}{32}} = 1.88mm$

In other words the relatively large radius of the tire in comparison to the asperity is such that the tire surface at the contact is approximately flat. Traditional fundamental rubber friction models propose friction to be composed of three aspects: adhesion, wear, and deformation. Adhesive friction has been shown to be a function of the critical shear stress and the plastic yield pressure

(Tabor, 1959), both of which are independent of macroscopic roughness and hence the increase in friction shown between the curves in Figure 7-8 must be explained by either increased deformation or increased wear.

Let us start by assuming that wear is generally independent of macroscopic asperity radius. From Greenwood and Tabor (1957), the friction force due to deformation is a function of the mean contact pressure, which from Hertzian theory is directly proportional to the peak contact pressure p_0, i.e.

$$F_{deformation} = f(p_0) \tag{7.14}$$

From Hertzian theory for line contacts, the contact semi-width can be calculated from Equation 7.15

$$a = 2\left(\frac{PR}{\pi E^*}\right)^{0.5} \tag{7.15}$$

It can also be stated that for a line contact the peak contact pressure is a function of the geometrical mean radius (R), specifically:

$$p_0 = f(R) = \left(\frac{WE^*}{\pi R}\right)^{0.5} \tag{7.16}$$

Knowing this, the contact pressure (p) at any point (x) inline with the surface from the centre of the contact can be calculated from Equation 7.17

$$p(x) = p_0\left(1 - \frac{x^2}{a^2}\right)^{0.5} \tag{7.17}$$

Assuming that the pressure acts normal to the interface, then the normal force on a small element (dS) would be equal to pdS (Figure 7-9), whilst the horizontal component of this force would thus be equal to:

$$pdS\sin\beta = pdS\,x/R \tag{7.18}$$

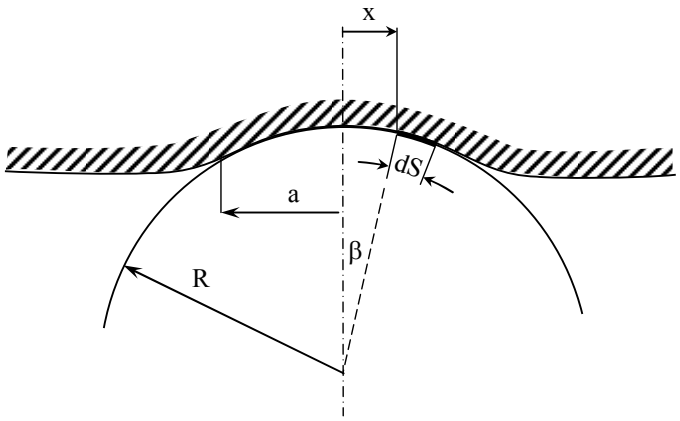


Figure 7-9: Small element in a deflected tire-surface interaction

For small values of deflection, as assumed by Hertzian theory, it can be stated for a unit-width of tire that $dS = dx$. Hence the total force produced by the sum of the elements becomes:

$$F = \int_0^a p\frac{x}{R}dx = \frac{p_0}{R}\int_0^a x\sqrt{1-\frac{x^2}{a^2}}dx = \frac{p_0 a^2}{6R} \tag{7.19}$$

Integrating:

$$F = \left(\frac{WE^*}{\pi R}\right)^{0.5}\left(\frac{WR}{\pi E^*}\right)\frac{4}{6R} \tag{7.20}$$

Simplifying:

$$F = \left(\frac{WE^*}{\pi R}\right)^{0.5}\frac{2W}{3\pi E^*} \tag{7.21}$$

Or in terms of friction coefficient μ:

$$\mu = \frac{F}{W} = \frac{2}{3\pi E^*}\left(\frac{WE^*}{\pi R}\right)^{0.5} = \left(\frac{W}{R}\right)^{0.5}\frac{2}{3\pi\sqrt{\pi E^*}} \tag{7.22}$$

Thus proving for a specific tire-surface combination:

$$\mu_{deformation} \propto W^{0.5} \tag{7.23}$$

And:

$$\mu_{deformation} \propto \left(\frac{1}{R}\right)^{0.5} \tag{7.24}$$

Assuming other variables remain constant, and ignoring the radius of the tire itself (valid since $R_{tire} >> R_{asperity}$), changing the radius of the asperity from 1mm to 2mm would lead to

approximately a 30% reduction ($\sqrt{1/2} \approx 0.7$) in the peak contact pressure and hence deformation friction.

If it is still assumed that wear is roughly independent of asperity radius, then an estimation can be made of the contribution of deformation to the total friction coefficient for this particular test case. Figure 7-8 shows that the peak friction coefficient for a radius of 2mm is approximately 80% of that for and an asperity radius of 1mm, indicating that for this particular surface and speed, deformation contributed to more than half of the total friction force.

Increasing the speed from 10mm/s to 20mm/s seems to reduce this contribution as shown in Figure 7-10.

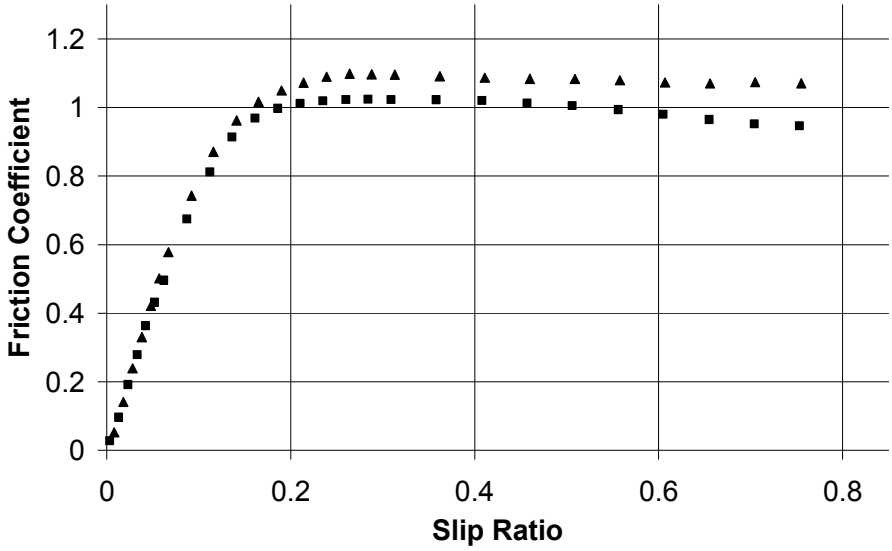


Figure 7-10: Effect of asperity radius at 20mm/s (▲=R1P2; ■ = R2P2)

The difference in percentage change between the two speeds (10mm/s and 20mm/s) implies that the contribution of each friction mechanism is not so rigid. It is likely that the smaller difference is due to an increase in friction due to wear. It is known that wear is a function of tire speed, which can be observed in 'tire marks' left by a vehicle coming to a stop. Initially the wear residues (rubber particles) are dense and darker, and as the vehicle's velocity lessens, the particulate density is reduced. Observation of the particles left on the cement surface confirmed this, although no formal wear particle measurements were taken.

The fact that the overall friction coefficient has only marginally increased implies that deformation too is a function of tire speed, which has not been researched in any great detail but

has been touched on in terms of reduced contact time, or the relaxation properties of the rubber compound (Myshkin, Petrokovets and Kovalev, 2005).

7.4.2.2 Effect of Pitch Change

Knowing that deformation friction is a function of peak pressure, Equation 7.23 shows that for a line contact, the contribution of deformation to the friction coefficient is proportional to the normal load per unit length of the tire. If it is assumed that there are *n* number of contacting points and that the normal load is approximately distributed evenly across them (which is not strictly true), then the total force produced by the deformation of each asperity can be shown to be:

$$\mu_{deformation} = nk\left(\frac{W}{n}\right)^{0.5} = k\frac{n}{n^{0.5}}W^{0.5} = kn^{0.5}W^{0.5} \tag{7.25}$$

This implies that if the number of contact points (asperities) is halved then the effect on deformation friction would be identical to doubling the asperity radius. Figure 7-11 shows the effect of doubling the pitch length from 2mm to 4mm, for an asperity radius of 1mm.

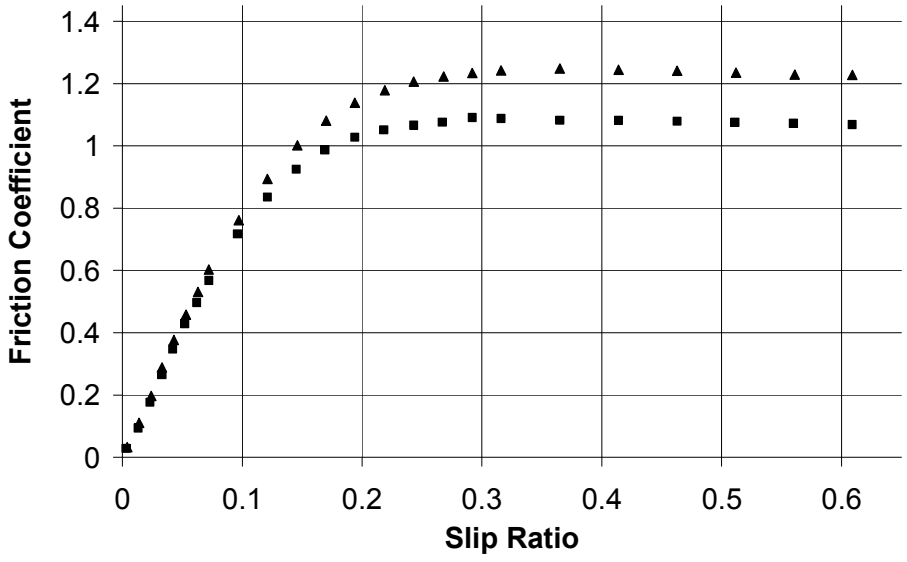


Figure 7-11: Effect of asperity pitch length at 10mm/s (▲=R1P4; ■ = R1P2)

Ink printing tests on the surfaces indicated that for a pitch length of 2mm, as shown in Figure 7-8, there are generally four individual asperities contacting the rubber surface at any time, whilst for a pitch of 4mm there are generally two contacting asperities. Figure 7-11 shows a similar reduction in friction coefficient to that found when doubling the radius. Given that the

other constants remained unchanged, this is exactly what would be expected, and indicates a similar contribution from deformation friction to the overall friction coefficient.

The same tests conducted at 20mm/s confirms the contribution of deformation friction shown previously (Figure 7-12).

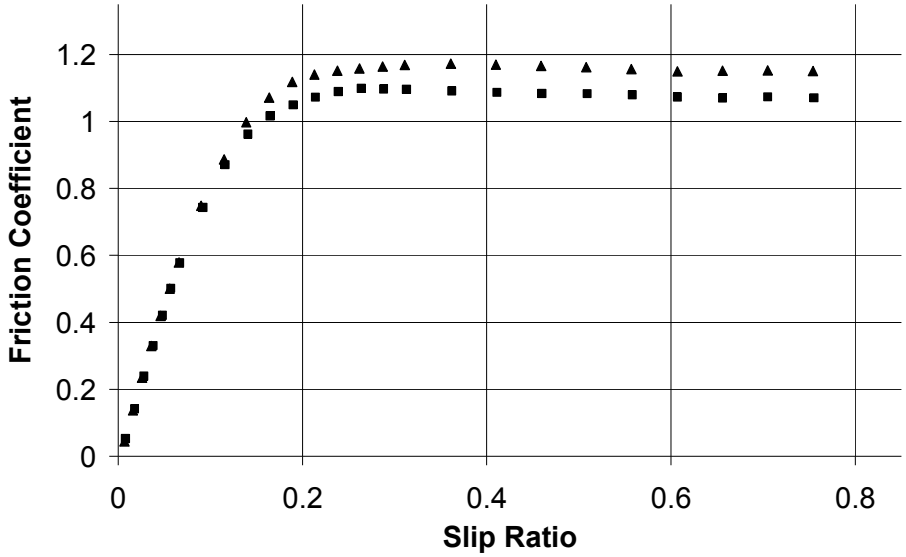


Figure 7-12: Effect of asperity pitch length at 20mm/s (▲=R1P4; ■ = R1P2)

7.5 Discussion

7.5.1 LuGre Model and Microscopic Roughness

Whilst a constant or point load distribution has been used thus far, a more accurate representation can be relatively easily incorporated into the LuGre model by modifying the load function. Previous research has offered several functions that offer more realistic distributions based on a flat surface. A typical normal load distribution during movement is shown in Figure 7-13 (taken from Gim and Nokravesh, 1999).

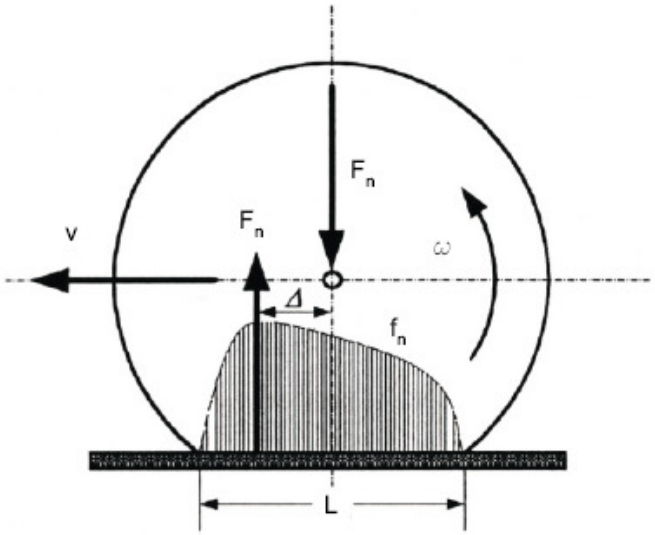


Figure 7-13: A typical load distribution pattern observed experimentally

This distribution pattern assumes an approximately smooth contacting surface as can be observed in microscopic roughness. In order to approximate this load distribution, Canudas-de-Wit et al. (2002) propose a combination of exponential and sinusoidal functions as shown in Equation 7.26.

$$F_n = k\int_0^L \exp(-\gamma\zeta)\sin(\pi\zeta / L)d\zeta \tag{7.26}$$

Where γ indicates the magnitude of the offset of the location of the maximum load point (a function of Δ in Figure 7-13).

Combining with Equation 7.10, the total friction force is given by:

$$F = k\int_0^L \left[g(v_r)\left(1-\exp\left(-\frac{\sigma_0}{g(v_r)}\left|\frac{v_r}{\omega r}\right|\zeta\right)\right)+\sigma_2 v_r \right]\exp(-\gamma\zeta)\sin(\pi\zeta / L)d\zeta \tag{7.27}$$

Integrating this leads to the rather untidy expression for total friction force shown in Equation 7.28.

$$k\left[\frac{e^{-\gamma L}L\pi\left(-g(v_r)\left(-c^2\left(1+e^{\gamma L}\right)L^2+2c\left(1+e^{\gamma L}\right)\gamma L+\left(-1+e^{cL}\right)\left(\gamma^2L^2+\pi\right)\right)+\sigma_2\left(1+e^{\gamma L}\right)\left((c-\gamma)^2L^2+\pi^2\right)v_r\right)}{\left((c-\gamma)^2L^2+\pi^2\right)\left(\gamma^2L^2+\pi^2\right)}\right] \tag{7.28}$$

Where:

$$k=\frac{\gamma^2L^2+\pi^2}{\pi L\left(e^{-\gamma L}+1\right)}$$

The value of gamma has no direct physical meaning, whilst the magnitude of gamma will vary depending on the other variables within the formula. Using the same values as in Table 27, the effect of γ changing gamma from 0 (dotted line) to 20 (solid line) on the load distribution is shown in Figure 7-14.

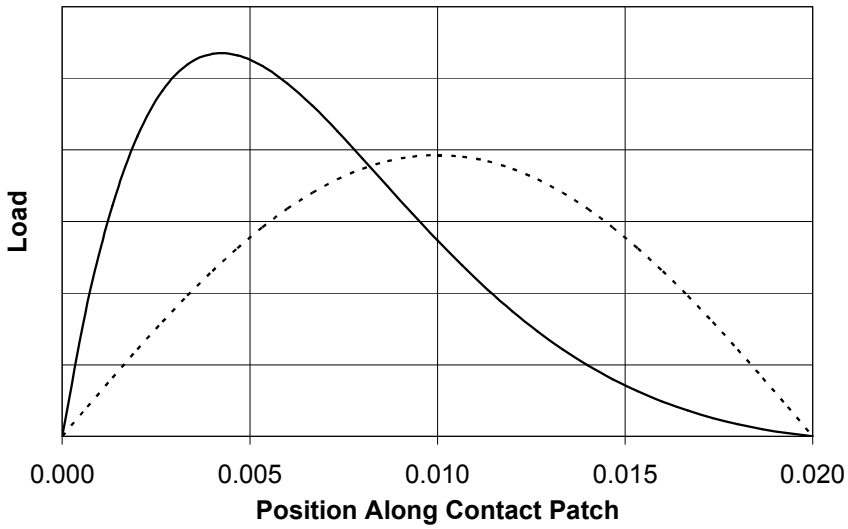


Figure 7-14: The Effect of gamma on load distribution

Although this leads to a higher 'peak' load, the total load (area under the curves) remains the same. It should be noted that for the case of braking, the value of γ must be positive, meaning that as the peak value of the load distribution moves away from the centre of the patch there is a reduction in the overall magnitude of the friction curve, as expected. This is shown in Figure 7-15.

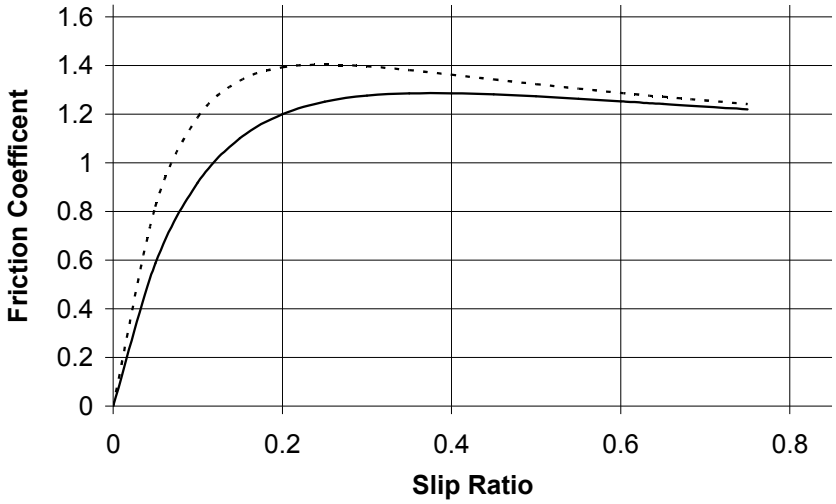


Figure 7-15: The effect of gamma on friction coefficient

This conclusion however is not entirely accurate; since it is reasonable to assume that the value of γ is not a constant, and instead is a function of the relative velocity (v_r), which itself is a function of both the slip ratio and tangential wheel speed. Hence it should perhaps be stated that instead of assuming that an experimentally obtained curve should precisely fit either of these curves, it should instead lay somewhere between them.

7.5.1.1 LuGre Model and Macroscopic Roughness

By manipulating the values of the friction coefficients in this equation it is possible to fit the LuGre curves to the experimental data with near perfect correlation, as shown in Figure 7-16.

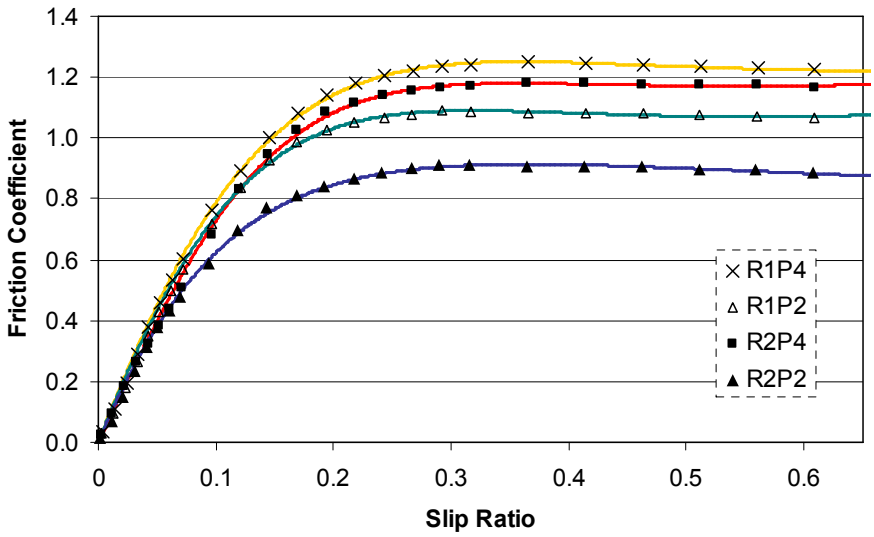


Figure 7-16: LuGre friction curves for macroscopic roughnesses

These curves could only be fitted by having a non-symmetric load distribution, as described by a sinusoidal-exponential load function. This was because the initial steepness of the curve could not be reproduced with a simple point or universal load distribution. This perhaps indicates that the basic LuGre model overestimates the initial tire friction available.

It is interesting to note that the order of the curves in terms of peak friction coefficient follow the same order as the predicted contact pressure, possibly indicating the relatively high importance of deformation friction under these friction conditions.

7.5.1.2 Tire Friction at Low Slip Values

Another interesting result that is observed when the four curves are shown on the same graph is highlighted in Figure 7-17, which figure shows the experimentally-measured friction coefficient at low slip ratios.

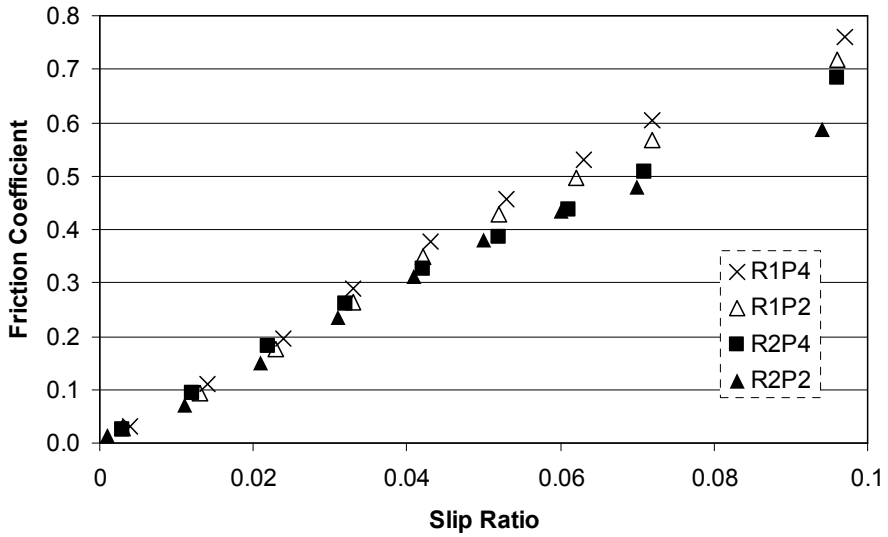


Figure 7-17: Macroscopic roughnesses friction curves at low slip values

Figure 7-17 shows that at higher slip ratios, the coefficient of friction is quite clearly larger when the pitch length is larger, irrespective of the radius of the asperities. Conversely at lower slip ratios the smaller asperity radius is more important in producing a higher friction. This finding suggests that the mechanisms that are involved in the friction coefficient are not quite as rigid as previously discussed. Given that ABS systems are designed to operate at lower slips ratios, near the peak friction coefficient, this result also suggests that in order to maximise the friction available, road surfaces should be constructed using smaller asperities, more sparsely distributed.

7.5.1.3 Immediate Variations in Friction Coefficient

One of the initial concerns with the test rig and the surface profiles was the immediate variations in friction coefficient that occur as the tire moves from one asperity to the next. In order to assess the importance of this effect it was necessary to eliminate macro changes in both the immediate friction force and vertical deflection, so the minor changes could be more easily observed. This was achieved by taking the difference from a moving average as the tire moved across the surface. The vertical deflection and immediate friction force was found to be largely in phase. This relationship is strongest in the profile that had a pitch of 4mm and an asperity radius of 1mm. This is shown in Figure 7-18.

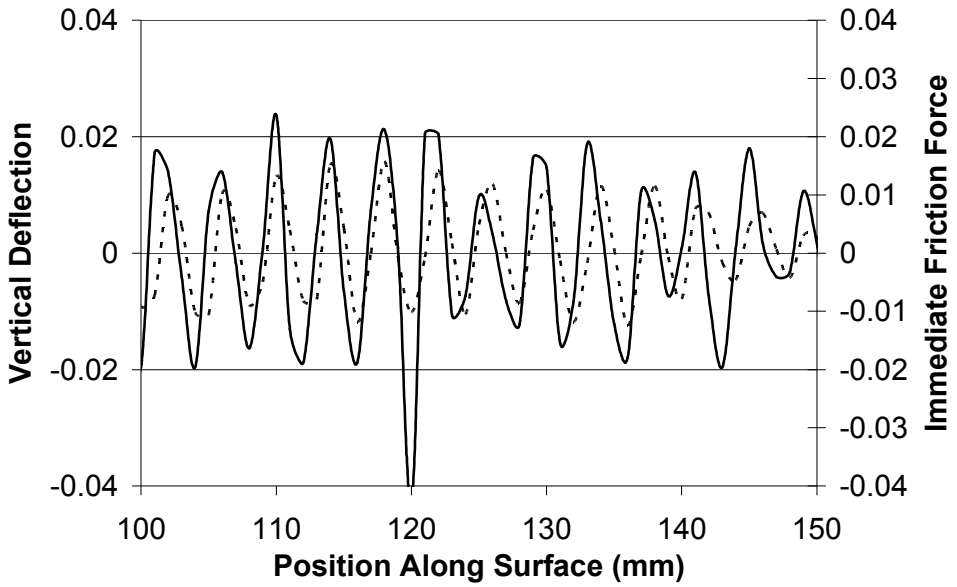


Figure 7-18: Vertical deflection (dotted) and immediate friction force (solid)

This effect could be due to the small amount of force that is lost as the tire has to climb out of each 'valley' between asperities. However, given that it has already been determined that the tire contacts a number of asperities simultaneously, this is unlikely. It is more likely that this result could be due to the change in load distribution that occurs as the tire moves across the asperities. If it is assumed that each sample's profile has measurably identical (or at least very similar) microscopic roughness, it can be stated that any variation in friction can only be attributed to changes in the actual (not apparent) contact length and contact pressure.

7.5.1.4 Multiple Contact Point Distribution

In order to improve the LuGre model's load distribution pattern, it needs to be adjusted to incorporate the possibility of multiple points of contact. Hence a new formula for the load distribution is presented that allows for the instantaneous changes in load distribution due to macroscopic roughness:

$$F_n = k\int_0^L \sin^2\left[\lambda + (\theta\pi\zeta / L)\right]\sin\left[\pi\zeta / L\right]\, d\zeta \tag{7.29}$$

Where the addition of λ and θ allows for multiple peaks at different points along the patch length (as a result of macroscopic roughness), whilst the second 'sine' term ensures that the normal load is always zero at the boundary of the contact patch. A typical example of this type of distribution is shown in Figure 7-19.

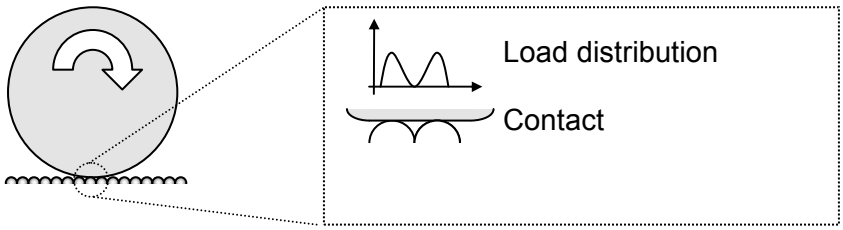


Figure 7-19: The instantaneous change in normal load distribution

From Equation 7.29 it can be seen that the value of λ can vary from 0 to π, which indicates the position of the asperities across the contact patch, whilst the value of θ indicates the number of asperities. Experimentation with these values shows that the instantaneous friction curve decreases significantly as the value of λ varies from 0 to $\pi/2$ and then rises again as λ approaches π. For a two asperity contact distribution ($\theta = 2$), the curve varies as shown in Figure 7-20 (the values of the constants are as shown in Table 27).

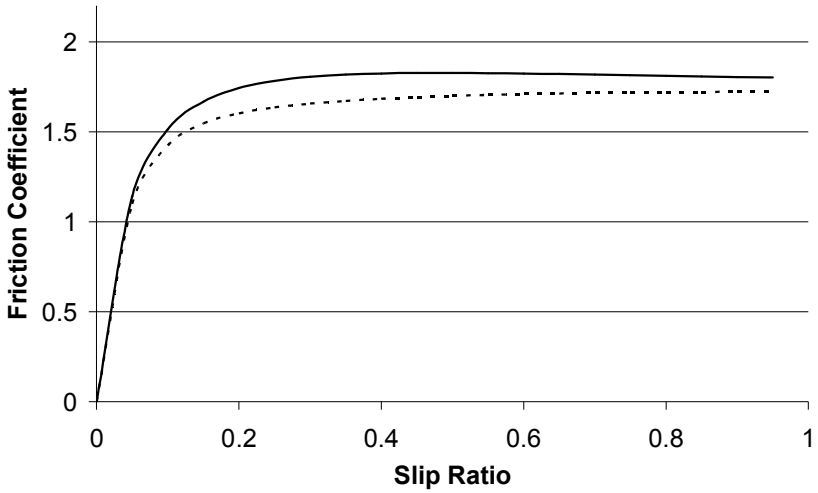


Figure 7-20: The change in immediate friction across two asperities

This implies that as a tire moves over macroscopic asperities the immediate friction can change considerably, simply because the position of the peak point of friction changes relative the axis of rotation, hence the moment changes. This assumes that the friction force is continuously applied parallel to macro surface, and hence does not take into account any vertical deflection of the tire due to these asperities.

Since it is extremely difficult to measure instantaneous friction over a wide range of values of slip ratio, friction is instead averaged across a constant roughness profile; hence the immediate friction becomes less important. When λ is averaged over its entire range, θ becomes irrelevant and the distribution returns to a simple sinusoidal shape, as shown in Figure 7-21. This figure shows a number of instantaneous load distribution curves (solid lines) together with an average load distribution curve (dotted line).

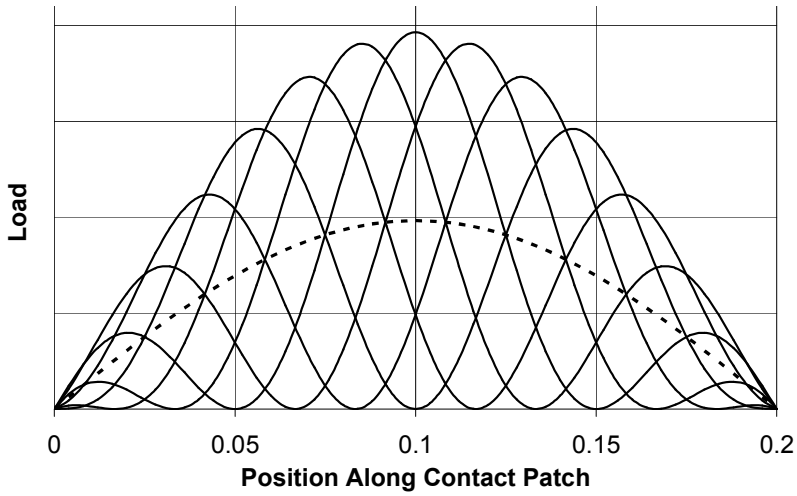


Figure 7-21: Average load distribution across multiple asperities

7.6 Conclusions

This complementary chapter has looked at existing tire friction models to assess the extent to which they can be used to represent physical phenomenon such as macroscopic roughness. It was found that traditional tire friction models (even those that claim to be based upon physical phenomena) are invariably only useful in simulation through 'trial and error' data fitting. Whilst they can be used to estimate the effect of load and velocity on friction, they cannot be used to predict the effect of physical phenomenon, and furthermore cannot be used to theoretically improve tire friction.

It was found that specific macroscopic roughness parameters such as asperity radius and pitch length have a very large influence on the deformation friction and hence the overall friction coefficient. At lower slip ratios it was found that a higher friction force can be achieved through using smaller radius asperities, whilst at higher slip ratios the pitch length has a stronger influence. This suggests that friction depends more on the contact pressure than the contact area, implying that the friction is governed more by the hysteresis (deformation) term than the adhesive term under the present contact conditions. Hence approximating the contact pressure using Hertzian theory can provide a very reasonable prediction of the overall effect of altering either the asperity radius or pitch length. The magnitude of the improvement in friction that can be obtained by careful selection of macroscopic roughness is comparable to using different microscopically rough surfaces.

The results shown have a number of implications for improving the representation of friction models towards real variations in road surfaces. Whilst these results cannot currently be used to improve vehicle safety systems, they do provide some useful insight into the effect of road surface conditions, which could possibly be used to improve tire friction in areas where artificial roughness is used to improve braking efficiency, such as when approaching traffic lights.

CHAPTER 8: COMPLEMENTARY RESEARCH: PROPOSED METHOD FOR CONTACT FILM THICKNESS MEASUREMENT

Parts of this chapter have been presented at the 2nd International Conference on Advanced Tribology and published in the associated proceedings.

8.1 Introduction

8.1.1 Chapter Summary

The behaviour and durability of traction fluids within traction drives has been well documented; however this behaviour is normally based around experimental test-rigs, which are not always capable of simulating real world situations. Given the relatively high costs of replacing traction drive parts due to the high tolerances required, failure of the fluid film can be disastrous. Even small changes in operating conditions beyond the design limits can quickly lead to wear and eventual failure, hence any system that can monitor the traction film thickness is potentially invaluable.

Film thickness measurements have been of great interest in the field of tribology for a number of years, primarily aimed at improving the lubrication of bearings. Some of these techniques can also be applied to traction fluid contacts, such as those within the CP-CVT. This chapter proposes the use of one particular method (electrical capacitance) that could potentially be used during the operation of the CP-CVT in order to constantly monitor the fluid film thickness to ensure it remains within a predetermined limit.

The proposed method monitors the total capacitance of all of the contacts, from which the film thickness can be extracted. In order to achieve this other contact parameters must be known, such as the contact size, which can be predicted from Hertzian theory and perhaps more importantly, the effect of pressure on the permittivity of the traction fluid. A specially modified test rig is used to determine this relationship, whilst a number of contact properties were derived, culminating in a proposed contact thickness measurement system.

8.1.2 Overview of Film Thickness Measurements

Electrical methods were proposed for the measurement of film thickness as far back as 1958 (Askwith, Cameron and Crouch, 1966) preceding later optical techniques that are still used to this day (Westlake and Cameron, 1967). Electrical methods for measuring the thickness of EHD films were largely replaced by the optical interferometry technique. This method, which was developed several decades ago is capable of achieving a resolution of less than five nanometres and is able to map the whole contact area, whilst electrical methods only provide overall information on the contact and are more susceptible to impurities within the lubricant. More recently, advances in technology and specifically in measurement systems have seen a resurgence in the interest of electrical methods, which are often seen as a better alternative since they allow the measurement of film thicknesses in metal on metal contacts, which are more representative of real lubricated contacts. For complex, in-situ measurements, such as is required in the CP-CVT, optical techniques are not a viable alternative as they require one surface material to be transparent. Despite this, optical methods have been studied in a much greater depth utilising various different techniques, and are considered sufficiently accurate today that they are used to calibrate other methods, such as capacitance. Other electrical methods, such as resistance or voltage discharge measurements, are generally seen as being less accurate than capacitance for very thin films, although they do have the advantage of having shorter measurement times.

The concept of capacitance for measurement is already routinely applied in automobiles to measure fluid levels such as petrol (Huber and Zorge, 1986). The sampling rate required from this application is significantly lower than what is required for film thickness measurements since the level of fluids within a vehicle generally do not change sharply during operation. Conversely for fluid film measurements, in order to ensure that even temporary film breakdowns are observed, samples need to be taken at much higher rate. Asperities contacting, or instantaneous metal-metal contact due to vibration can last less than a few milliseconds, but can have devastating effects on the life of a traction drive. The concept of film thickness measurements using capacitance is based around the theory of parallel plate capacitors, which states:

$$C = \varepsilon_r \varepsilon_0 \frac{A}{d} \tag{8.1}$$

Where: ε_r is the relative static permittivity of the intermediary material, ε_0 is the electric constant (8.854×10^{-12}), A is the area of the plates, and d is the distance between them. Hence if it is assumed that the permittivity and area are known, then the gap between the two objects can be determined. It is this theory that is used to determine film thickness.

8.2 Methodology

8.2.1 EHD Contact Analysis

In order to reasonably predict the film thickness of a contact based on capacitance alone, a number of factors need to be determined, such as the contact dimensions (area), the permittivity of the traction fluid at increased pressures, and the effect of entrainment speed on film thickness. The contact dimensions can be reasonably predicted from Hertzian theory, whilst the film thickness can be determined from the work of Chittenden et al. (1985), or approximated from the work of Anghel, Glovnea, and Spikes, (2004). The main difficulty is determining the effect of contact pressure on permittivity, which is critical to estimating accurate film thicknesses based on capacitance. The monitoring system proposed is designed to compare the film thickness derived from a capacitance measurement with what the film thickness should be based on theory. Both of these calculations require a number of additional measurements, as discussed. The proposed measurements required are shown in Figure 8-1.

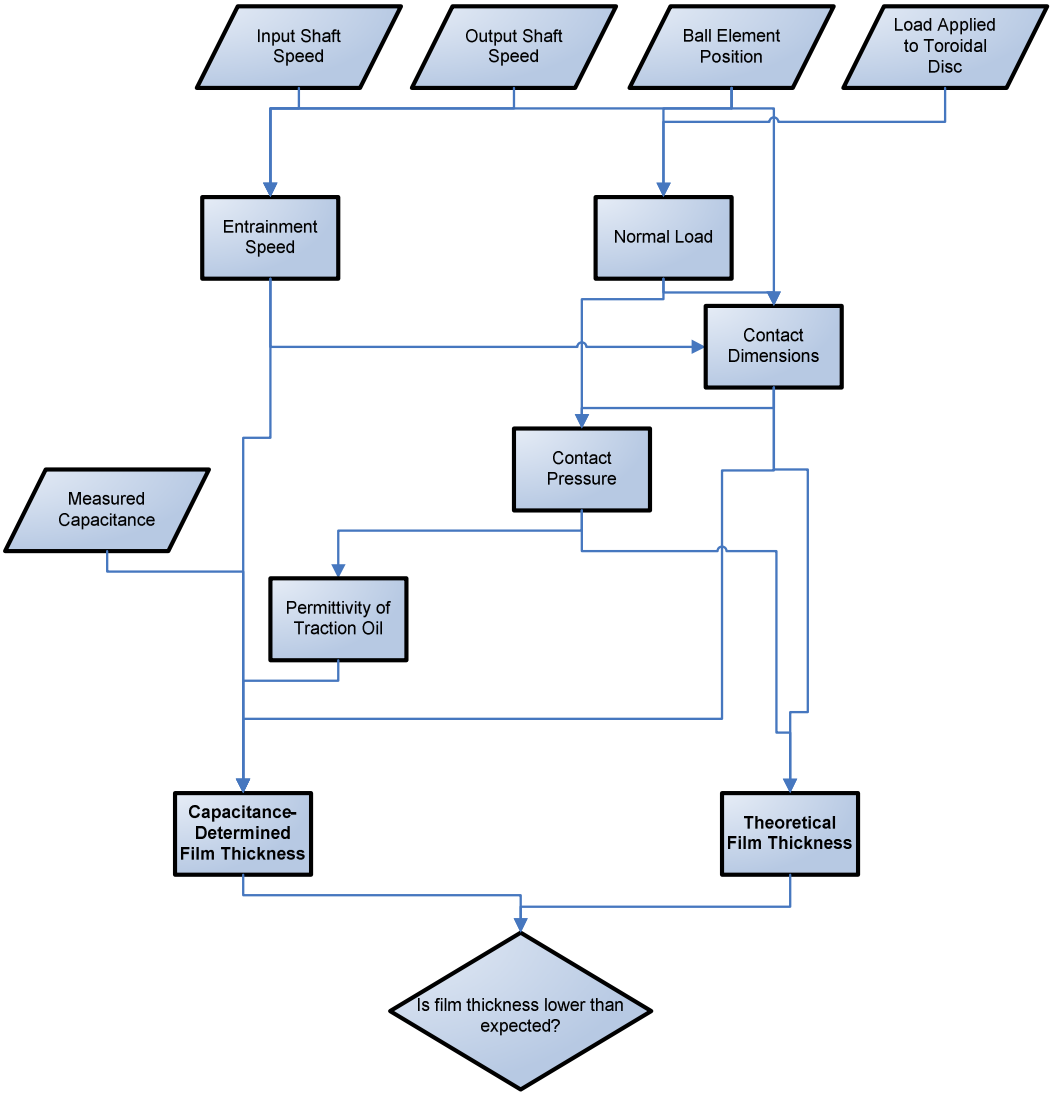


Figure 8-1: Summary of film thickness calculation processes

8.2.1.1 Contact Dimensions

The contact equations originally developed by Hertz (Johnson, 1985) are widely accepted as being an accurate model for the predictions of the contact dimensions. Since every contact within the CP-CVT has a surface which is non-uniform in two axes, each contact will be elliptical. The actual dimensions of the ellipse (*a* and *b*) are not entirely relevant since it is the contact area that is of interest, hence the contact area is simply:

$$A = \pi c^2 \tag{8.2}$$

Where:

$$c = \left(\frac{3}{4} \frac{W R_\sigma}{\overline{E}} \right)^{1/3} F_1 \tag{8.3}$$

Where W is the normal load, R is the reduced radii of curvature (which can be calculated from Equation 8.4), and $\overline{E}$ is the reduced contact modulus, shown in Equation 8.5.

$$R_\sigma = \sqrt{R_x R_y} \tag{8.4}$$

Where:

$$\frac{1}{R_x} = \frac{1}{R_{x1}} + \frac{1}{R_{x2}} \quad \text{And} \quad \frac{1}{R_y} = \frac{1}{R_{y1}} + \frac{1}{R_{y2}}$$

And:

$$\overline{E} = \left(\frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2} \right)^{-1} \tag{8.5}$$

It was determined earlier that the value of F_1 can be approximated as:

$$F_1 = -0.18 \log\left(R_x / R_y \right)^{1/2} \tag{8.6}$$

8.2.1.2 Contact Thickness and Pressure

Although the contact thickness can be estimated using the theory of parallel capacitance, it is also necessary to estimate the film thickness based on theory to determine what it should be, and whether the current film thickness is acceptable. The contact thickness is relatively difficult to determine, since it is affected by a number of different parameters. A good model for film thickness prediction was proposed by Chittenden et al. (1985), whom stated:

$$\frac{h_c}{R_\sigma} = 3.0 \left(\frac{U\eta}{\overline{E} R_\sigma} \right)^{0.67} \left(\alpha \overline{E} \right)^{0.49} \left(\frac{W}{\overline{E} R_\sigma^{\,2}} \right)^{-0.073} \tag{8.7}$$

In addition to film thickness it is necessary to determine the contact pressure since it will influence the permittivity of the traction oil, and hence the measured capacitance. The contact pressure can be estimated from Hertzian theory, which states the peak pressure is:

$$p_0 = \left(\frac{1}{\pi}\right)\left(\frac{6W\bar{E}^2}{R_\sigma{}^2}\right)^{1/3} F_1^{-2} \tag{8.8}$$

Whilst the mean contact pressure is simply:

$$\bar{p} = \frac{P}{\pi ab} = \tfrac{2}{3} p_0 \tag{8.9}$$

These two models for contact thickness and pressure are simplified and do not take into account variation across the contact, which is shown in Figure 8-2 (Stachowiak and Batchelor, 2005)

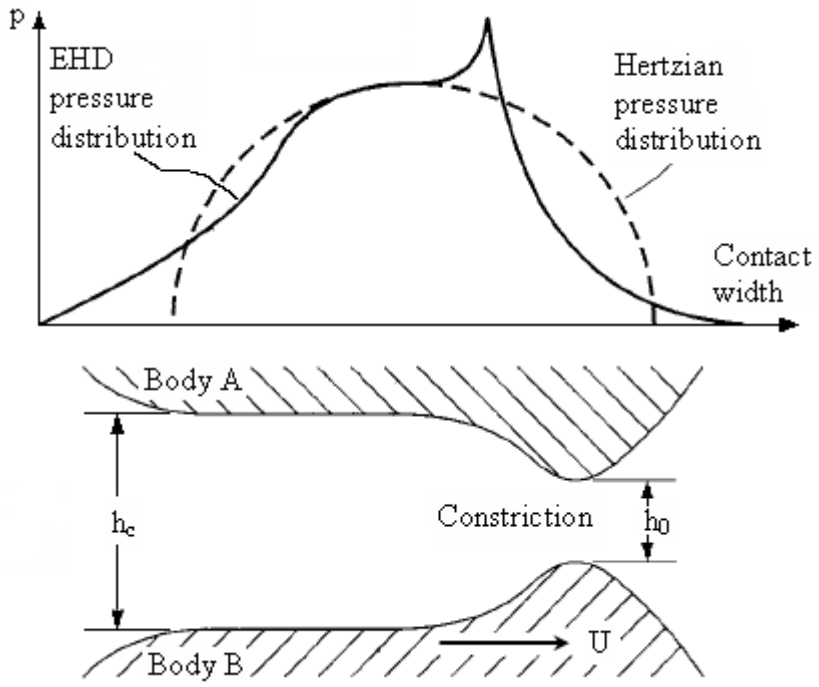


Figure 8-2: Hydrodynamic pressure distribution in an elastohydrodynamic contact

This figure shows that both the film thickness and contact pressure vary significantly over the width of the contact. Furthermore, since both of these factors significantly affect capacitance, the contact capacitance will vary considerably across the contact as well. By overlaying film thickness and contact pressure variations, it is possible to determine the approximate capacitance variation across the contact, as shown in Figure 8-3.

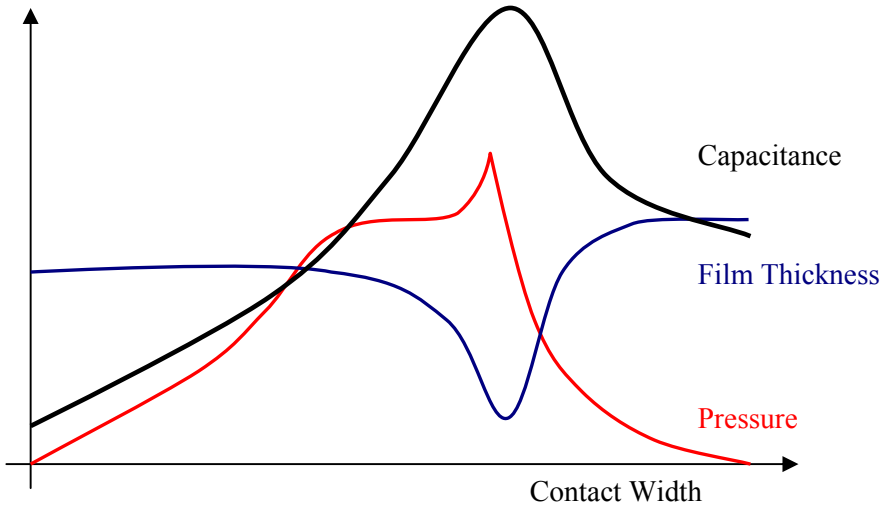


Figure 8-3: Capacitance variation across contact width

This variation of capacitance across the contact is too complex to incorporate into predictions of film thickness, and hence needs to be simplified. This can be achieved by using a similar approach to Cameron (1985), whom considered that the film is completely flat and that the shape of the deformed surfaces in the inlet of the contact is given by Hertzian theory, as shown in Figure 8-4.

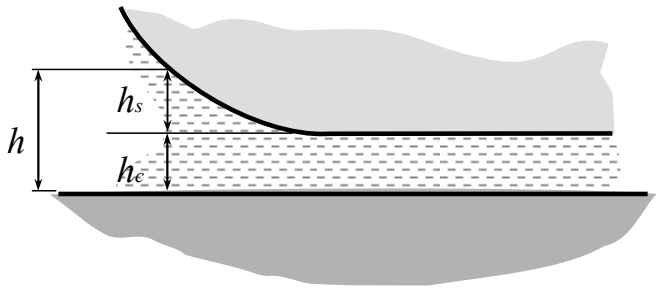


Figure 8-4: Cameron's simplified film thickness model

This simplified model of film thickness allows the contact to be approximated as a straightforward parallel plate capacitor, with the area (A) equal to the area of the contact and the plate separation (d) to be approximated as the central film thickness h_c.

8.2.1.3 Influence of Capacitance outside of Contact Area

In addition to the contact capacitance, there will be an additional capacitance between the ball elements and discs that occur because of the air gap surrounding the contact. Given that the permittivity of air is less than half that of oil and since the distance between the surfaces is

larger; the influence of this additional capacitance is theoretically very small. To assess the contribution of this capacitance, we will begin by ignoring the curvature of the discs, simplifying the problem to a ball-on-flat arrangement, as shown in Figure 8-5.

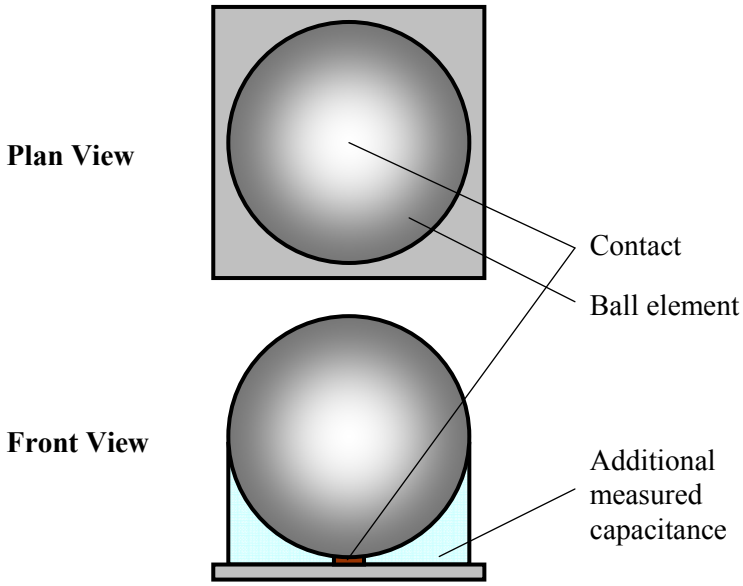


Figure 8-5: Approximated ball-on-flat capacitance model

To assess the capacitance of this arrangement and hence the contribution of capacitance outside of the contact, Figure 8-6 shows approximate capacitance field lines, which are assumed to run perpendicular to each surface (Hudlet et al. 1998)

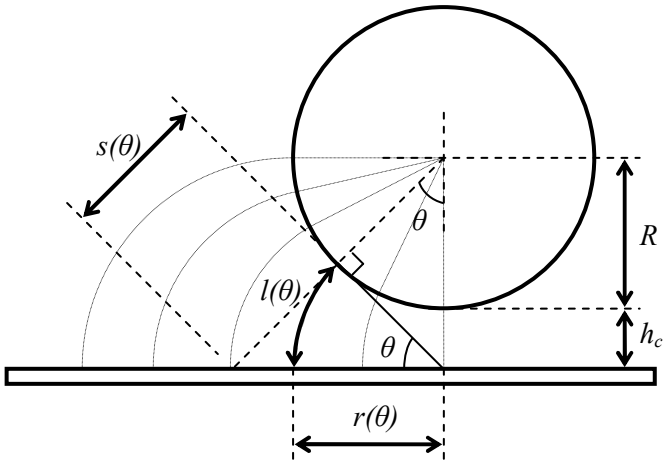


Figure 8-6: Approximate capacitance field lines of ball-on-flat arrangement

If infinitesimally small field line is taken as an individual capacitor, then the total capacitance will be given by the summation of each of these capacitors. Assuming that the field lines are identical around the circumference of the ball element, then the area of each field line of unit width $Rd\theta$ will simply be:

$$dA = 2\pi R \sin\theta \times Rd\theta = 2\pi R^2 \sin\theta d\theta \tag{8.10}$$

Whilst the total capacitance will be (adapted from Hudlet et al., 1998):

$$C = \varepsilon_r \varepsilon_0 \int_0^{\pi/2} \frac{A(\theta)}{l(\theta)} d\theta \tag{8.11}$$

The length of each field line, $l(\theta)$ can be calculated from simple geometry. Hence if

$$s(\theta) = \frac{h_c + R}{\cos\theta} - R \tag{8.12}$$

Then:

$$r(\theta) = \frac{\frac{h_c + R}{\cos\theta} - R}{\tan\theta}$$

Or:

$$r(\theta) = \frac{(h_c + R)\cos\theta}{\cos\theta \sin\theta} - \frac{R\cos\theta}{\sin\theta}$$

Simplifying:

$$r(\theta) = \frac{(h_c + R) - R\cos\theta}{\sin\theta} \tag{8.13}$$

And hence:

$$l(\theta) = \frac{\theta[h_c + R(1-\cos\theta)]}{\sin\theta} \tag{8.14}$$

Hence from Equation 8.11 and assuming that the upper half of the sphere has negligible influence on total capacitance:

$$C = \varepsilon_r \varepsilon_0 \int_0^{\pi/2} \frac{2\pi R^2 \sin\theta d\theta}{\frac{\theta[h_c + R(1-\cos\theta)]}{\sin\theta}} = 2\pi R^2 \varepsilon_r \varepsilon_0 \int_0^{\pi/2} \frac{\sin^2\theta}{\theta[h_c + R(1-\cos\theta)]} d\theta \tag{8.15}$$

However, what is of interest is the contribution of capacitance that occurs outside of the contact. For simplicity, we will assume that contact is approximately circular and hence the radius of the contact is simply c, which can be calculated from Hertzian theory. Hence we require two

separate integrals; one for the capacitance inside the contact (where $\varepsilon_r = \varepsilon_{oil}$), and one for the capacitance outside of the contact (where $\varepsilon_r = \varepsilon_{air} \approx 1$). Hence:

$$C = 2\pi R^2 \varepsilon_0 \left[\varepsilon_{oil} \int_0^{\arcsin(c/R)} \frac{\sin^2\theta}{\theta[h_c + R(1-\cos\theta)]} d\theta + \int_{\arcsin(c/R)}^{\pi/2} \frac{\sin^2\theta}{\theta[h_c + R(1-\cos\theta)]} d\theta \right] \tag{8.16}$$

Using Equation 8.16 it is now possible to assess the contribution of the region outside of the contact to the total capacitance. This will obviously vary depending on factors such as the radius of the ball element, the lubricant temperature, and more importantly the film thickness, as shown in Figure 8-7, which assumed a ball radius of 30mm and an oil permittivity of 2.2.

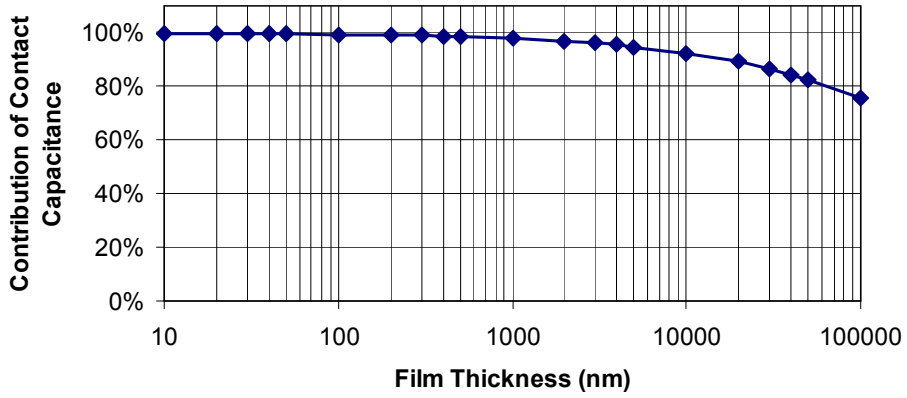


Figure 8-7: Contribution of contact capacitance to overall ball-on-flat capacitance

This Figure shows that for film thicknesses smaller than 1000nm (1μm), the capacitance of the contact dominates the measured capacitance under these conditions. Given that it is small film thicknesses that are of most concern; it can be reasonably assumed that the capacitance outside of the contact is negligible for a simplified ball on flat arrangement. This effect is exaggerated further since due to surface deformation, the contact is closer to parallel than assumed by this model. Replacing the sphere-on-flat arrangement with the simpler parallel plate model increases the capacitance of the contact by up to a factor of 5.

It is therefore reasonable to assume that the film thickness can be extracted directly from the capacitance measured (C) as follows:

$$h_c = \frac{\pi c^2}{C} \varepsilon_0 \varepsilon_{oil} \tag{8.17}$$

The only unknown in the equation is the permittivity of the traction oil, which varies with contact pressure and lubricant temperature.

8.2.2 Permittivity of Traction Fluid

To determine the film thickness of the contacts within the CP-CVT during operation it is necessary to determine the variation of the traction oil permittivity with temperature and pressure. This has been studied previously by Gilchrist et al. (1957) who looked at the effect of pressure on the static-dielectric (frequency-independent) permittivity of 1-propanol and glycerol. They found that the permittivity increased with pressure, albeit less than predicted from the relative increase in fluid density. More recently, Vij et al. (1978) attempted to indentify changes in the molecular structure of heptanol-isomers and specifically the variation of the relative permittivity with temperature and pressure. They found that the electric permittivity decreased with temperature and increased with pressure, supporting what was found earlier by Gilchrist et al. (1957). Hence it can be reasonably assumed that the permittivity of oil decreases with temperature and increases with the pressure. In an elastohydrodynamic contact the lubricant is subjected to a large pressure pulse and temperature rise due to inlet compression and shear, however the studies discussed do not generally approach the magnitude of pressure change found in EHD and traction contacts. In order to determine the behaviour of the lubricant in EHD environment it is necessary to measure the capacitance of a contact in experimentally identical conditions.

Initially tests were conducted using a simple parallel plate capacitor to determine the permittivity of two different lubricant oils (polyalphaolefin (PAO) and a mineral oil) at standard atmospheric pressure and temperature. These oils, although fundamentally different from traction fluids are presumed to demonstrate a similar pressure-permittivity dependency. At standard atmospheric pressure and temperature the static permittivity of the oils was found to be 2.2.

In order to measure the static-permittivity of the oils in EHD conditions, a standard PCS Instruments optical interferometry test rig was modified to allow the collection of an electrical signal from both the ball carriage and the contacting disc, as shown in Figure 8-8.

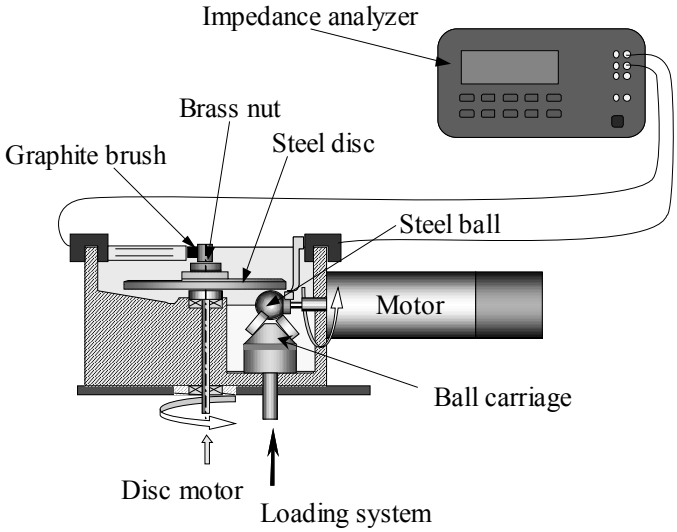


Figure 8-8: Modified PCS Instruments test rig to allow capacitance measurements

The contact itself was formed between a steel disc and steel ball, both with mirror-like surface finishes. Both the disc and ball were driven independently to allow variation in the entrainment speed and the slide-roll ratio.

8.3 Results of Permittivity Tests

Initially an optical technique (ultra-thin film interferometry; UTFI) was used to determine the contact thickness at a variety of entrainment speeds, as shown in Figure 8-9. These tests were conducted at 25°C with a normal load of 10N.

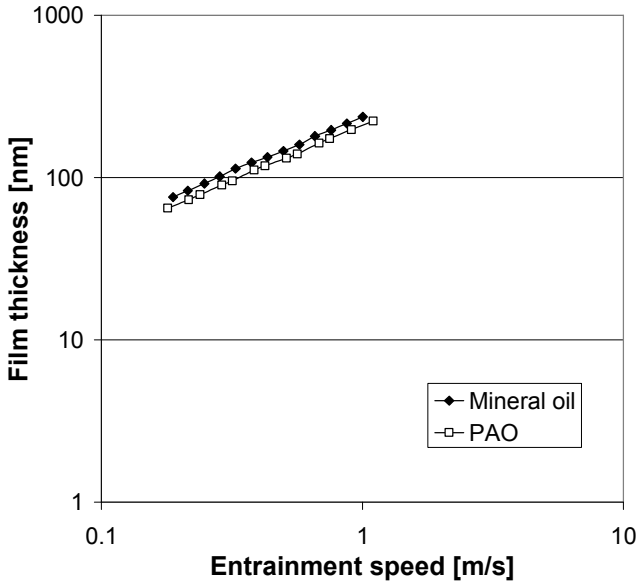


Figure 8-9: UTFI-obtained film thickness as a function of entrainment speed

Once the film thickness had been determined as a function of entrainment speed, tests were carried out using the steel disc and the capacitance of the contact was measured using a phase-gain analyser. In order to take into account the larger elastic modulus of the steel disc in comparison to glass, the tests were conducted with a normal load of 20N, rather than the 10N used previously. The results of the capacitance tests are shown in Figure 8-10

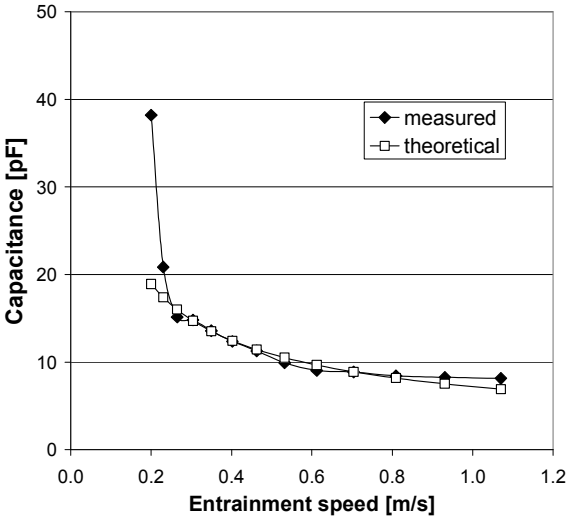


Figure 8-10: Effect of entrainment speed on measured and theoretical contact capacitance

This figure (which has been corrected for the influence of background capacitance) shows a comparison of the measured capacitance with the equivalent capacitance expected assuming the contact can be approximated as a parallel plate. In order to obtain the correlation shown, the relative permittivity must have increased by a factor of 3. It is plausible that this increase could be a result of the relatively large contact pressure, or alternatively because of the simplification of the contact as a parallel plate, which as discussed earlier, is only an approximation. This is supported by the larger deviation found at lower entrainments speeds, where the film thickness is smaller and hence the parallel-plate approximation will have a greater influence.

Another possibility is that the larger values of measured capacitance are simply a result of the influence of the region surrounding the contact, which at larger film thicknesses has more of an influence as shown in Figure 8-7. It is possible therefore that the correction made for background capacitance was not large enough, meaning that the actual permittivity is lower than calculated.

8.4 Proposed Design of Monitoring System

8.4.1 Collection Point and Circuit Layout

In reality, the CP-CVT does not consist of a single contact-capacitor, but instead consists of a complex arrangement of different capacitors acting both in series and parallel, as in Figure 8-11.

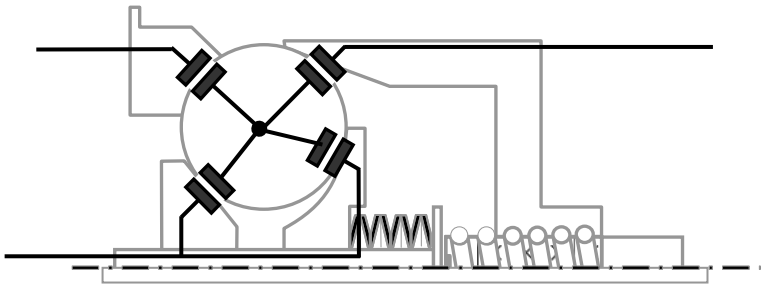


Figure 8-11: Contact capacitance within the CP-CVT

This entire capacitor arrangement is in parallel with two or three identical arrangements representing each of the ball elements in the cage. If a signal is simply collected at the input and output shafts then the determination of each contact capacitance becomes much more difficult. An alternative arrangement is proposed, in which a signal is applied to the ball separator, whilst both the input and output shafts are grounded, as shown in Figure 8-12.

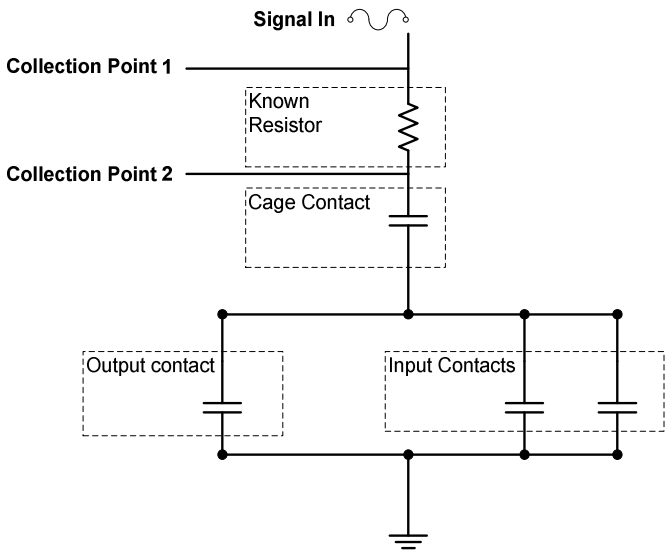


Figure 8-12: Proposed layout of capacitance monitoring system

This arrangement, which essentially produces an impedance 'ladder', ensures that each of the critical contacts (conical/toroidal input disc and conical output discs) is in parallel. This is important since capacitors in parallel can simply be added, as shown in Equation 8.18.

$$C_{total} = C_1 + C_2 + C_3 + \ldots \tag{8.18}$$

The reason this is advantageous is that the measured capacitance will be dominated by the largest contact capacitance. Hence as a contact's film thickness decreases, or a contact begins to become starved of lubricant, its capacitance will increase sharply, which will be noticeable immediately irrespective of the capacitance of the other contacts.

Additionally, by using the arrangement shown in Figure 8-12, the input and output shafts, which maybe subjected to additional interference from the engine, etc, are simply grounded, potentially eliminating interference at these points.

8.4.2 Methods of Capacitance Measurement

8.4.2.1 Capacitance Discharge

Typically, capacitance measurement circuits either use the principle of an impedance bridge or capacitance discharge. The latter uses the theory of either charge or discharge of a capacitor in an *RC* (resistor-capacitor) circuit. The basic equation is simply:

$$t = -RC\ln\left(1-\frac{V}{V_{in}}\right) \tag{8.19}$$

Thus, if the resistance (R) and voltage applied (V_{in}) are known, then the capacitance (C) can be determined by measuring the time it takes for the voltage across the capacitor (V) to reach a predetermined value. This approach is good for measuring fixed values of capacitance since it is relatively simple, and does not require any complicated circuitry. However this is only valid for fixed capacitance values and cannot be reasonable applied to situations where the capacitance changes with respect to time.

8.4.2.2 Schering Impedance Bridge

Alternatively, an impedance bridge can be used to determine capacitance. This arrangement usually takes the form of a modified Wheatstone bridge, known as a Schering bridge, as shown in Figure 8-13 (Gilson, 1998).

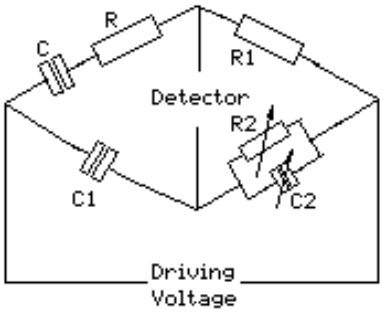


Figure 8-13: A typical Schering bridge

This type of bridge uses a variable capacitor and resistor (*C2*, *R2*) to balance the signal across two impedance ladders. Once the signals are balanced, the values of the unknown capacitor and resistor (*C* and *R* respectively) can be calculated from the variable capacitor and resistor values, as shown in Equations 8.20 and 8.21.

$$C = C1 \times (R2 / R1) \tag{8.20}$$
$$R = R1 \times (C2 / C1) \tag{8.21}$$

The advantage of this type of circuit is that only one of the variable components appears in each equation, and furthermore, it is independent of frequency (Gilson, 1998). Since the unknown capacitance *C* is normally the desired value, it can be accurately determined using only a variable resistor, which is a standard electrical component, the value of which can be precisely found. The adjustment of the variable resistor/capacitor is normally done manually for most applications, and hence it is again difficult to apply this type of circuit to measure a dynamically changing capacitor.

8.4.2.3 Simplified Impedance Bridge

Although a Schering Bridge or capacitance discharge method will give more accurate results, they require at least several seconds to measure single capacitance values. A simpler, albeit less accurate circuit simply uses a series *RC* circuit to produce an impedance ladder, as shown in Figure 8-14.

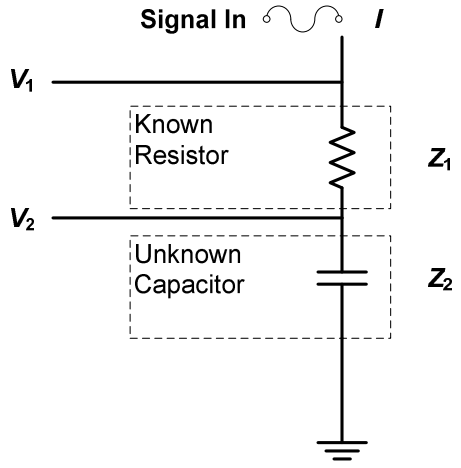


Figure 8-14: Simple *RC* series circuit

In this simple circuit, the voltage at each collection point is a function of the current applied and the impedance of the circuit:

$$V_1 = IZ_T$$
$$V_2 = IZ_2$$

Combining:

$$\frac{V_2}{V_1} = \frac{IZ_2}{IZ_T} = \frac{Z_2}{Z_2 + Z_1} \tag{8.22}$$

Comparing Figure 8-14 to Figure 8-12, it can be seen that impedance Z_2 is due to the combined capacitance of the contacts. If the resistance of the contact is ignored, then this impedance is purely due to the reactance of the contact ($-jX_c$), whilst Z_1 is only the resistance R, hence:

$$\frac{V_2}{V_1} = \frac{-jX_c}{-jX_c + R} = \frac{-j}{-j + (R/X_c)} \tag{8.23}$$

So the magnitude is simply:

$$\left|\frac{V_2}{V_1}\right| = \frac{1}{\sqrt{1 + (R/X_c)^2}} \tag{8.24}$$

Rearranging:

$$R/X_c = \sqrt{\left|\frac{V_1}{V_2}\right|^2 - 1} \tag{8.25}$$

X_c is simply the reactance due to the capacitance at a particular frequency, and can be calculated using Equation 8.26:

$$X_c = \frac{1}{2\pi f C_T} \tag{8.26}$$

Combining Equations 8.25 and 8.26:

$$C_T = \frac{\left(\sqrt{|V_1/V_2|^2 - 1}\right)}{2\pi f R} \tag{8.27}$$

From Figure 8-12, the total capacitance in the arrangement shown inside the CP-CVT is:

$$C_T = \frac{n}{(C_{co} + C_{ci} + C_{ti})^{-1} + C_{ca}^{-1}} \tag{8.28}$$

Where the subscripts *co*, *ci*, *ti* and *ca* indicate the capacitance between a single ball element and the conical input disc, conical output disc, toroidal input disc, and cage respectively.

8.4.3 Predicted Results of Capacitance Monitoring System

Using the Equations stated, the individual capacitance of each contact can be determined. It is assumed here that the permittivity of the oil approximately doubles at the contact pressure, and that the contacts themselves can be approximated as parallel plates of area πc^2, where the value of c for each contact is taken as calculated previously in Appendix A1. The film thickness of each contact is also taken from the previously calculated values. By reducing these thicknesses by a particular percentage it is possible to monitor the effect on the predicted capacitance of the contact between a single ball elements and each disc, as shown in Figure 8-15.

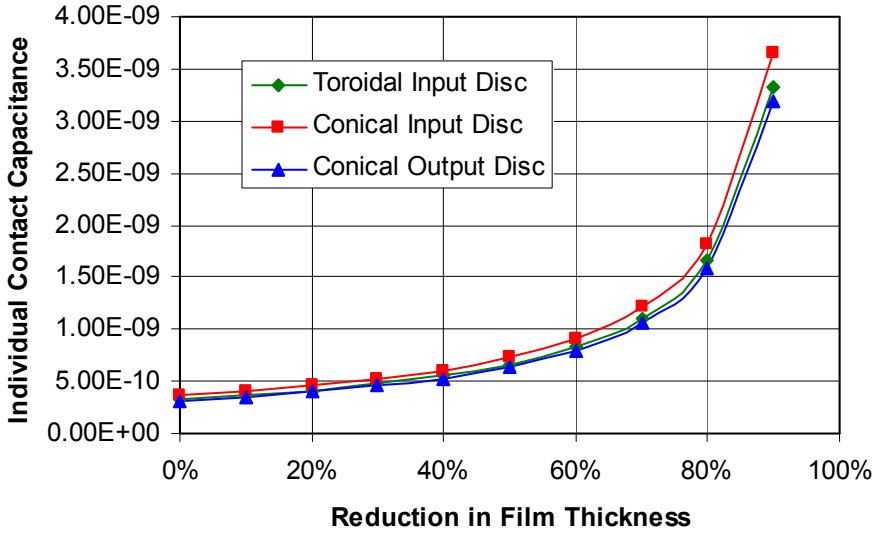


Figure 8-15: Effect of reduction in film thickness on individual contact capacitance

As expected, a reduction in the film thickness (possibly indicating contact failure) is shown by a significant increase in the individual contact capacitance. However what is of interest is how this translates to a change in the total measured capacitance. If it is assumed that only one contact fails, Equation 8.28 becomes significantly more complicated. If, for example one of the conical output contact capacitances changes to a new value C_{co}^*, then the total capacitance now becomes:

$$C_T = \frac{1}{\left[n(C_{ci} + C_{ti}) + ((n-1)C_{co}) + C_{co}^*\right]^{-1} + (nC_{ca})^{-1}} \tag{8.29}$$

Using this Equation and the capacitance values shown in Figure 8-15, the effect of a reduction in film thickness of an individual contact on total measured capacitance can now be determined. This is shown in Figure 8-16, which assumes four ball elements.

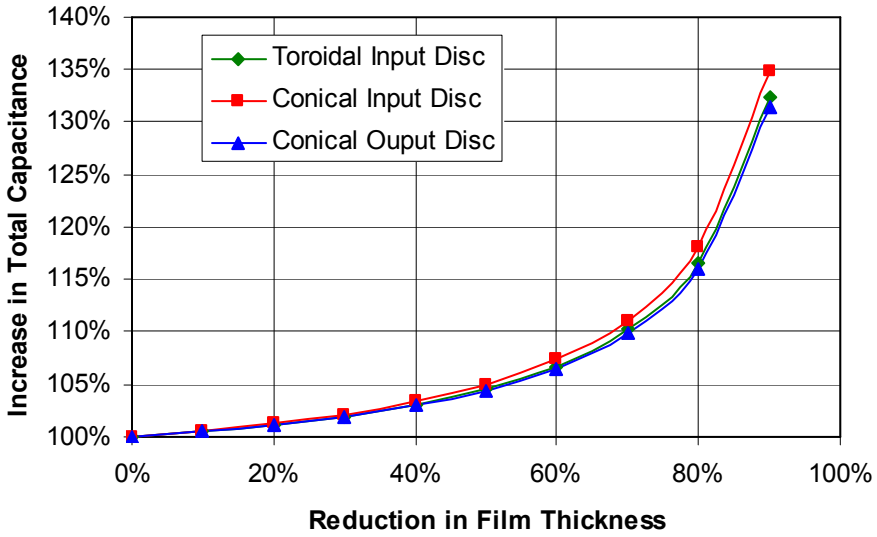


Figure 8-16: Effect of reduction in film thickness on total measured capacitance

This figure shows that even if only one of the 12 possible contacts begins to fail, the effect is immediately shown in the total measured capacitance, highlighting the advantage of using the parallel arrangement shown in Figure 8-12.

Furthermore, using the circuit shown and Equations 8.24 and 8.25, it is possible to predict how a reduction in film thickness of one contact will affect the measured parameter V_2/V_1, as shown in Figure 8-17, which assumes a frequency of 5kHz and a fixed resistor of 50kΩ.

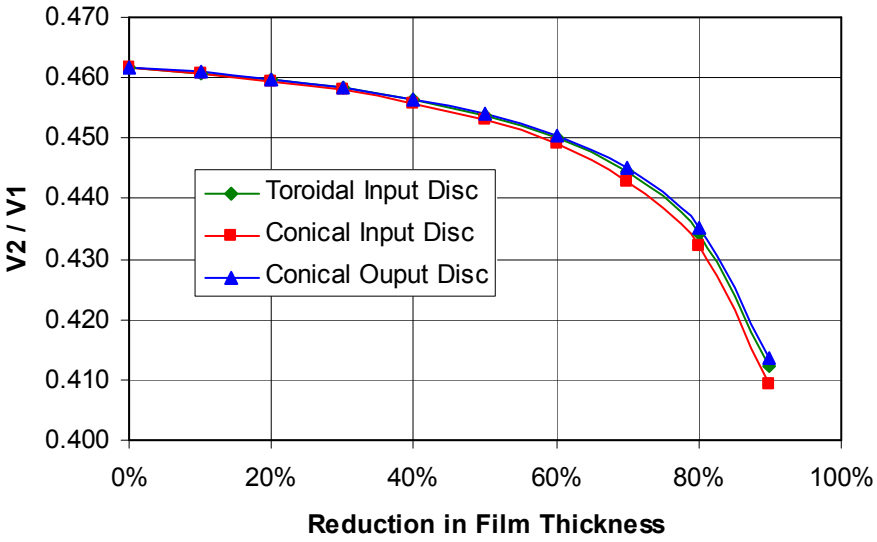


Figure 8-17: Effect of reduction in film thickness on measured value V_2/V_1

8.5 Conclusions

This chapter has presented a proposed method for the measurement and monitoring of the film thickness of the EHD contacts within the CP-CVT. The EHD contacts within traction drives are crucial to their operation, and any failure of the lubricating film can be catastrophic, leading to significantly increased wear, or even failure.

The proposed method utilises the theory of parallel-plate capacitance, which states that the distance separating the plates (the film thickness in this case) is inversely proportional to the capacitance measured. Hence if the film thickness decreases due to abnormal temperature rise or lubricant starvation, the capacitance across the contact would increase substantially. To determine the magnitude of the film thickness change, the contact area and permittivity of the lubricant must also be known. The proposed method uses Hertzian theory to determine the contact area, whilst the permittivity of the oil was determined by a series of capacitance tests conducted on a modified PCS Instruments test rig. The test rig measured the capacitance of an EHD contact formed between a steel disc and a steel ball, which was then compared to film thickness results found using a known optical technique in identical EHD conditions. These tests showed that the permittivity of the oil increases by a factor of approximately 3 due to the increased contact pressure.

The method proposed for measuring the capacitance of the CP-CVT contacts (and hence film thickness) is based on a simple impedance ladder, where an additional impedance of known value is introduced to determine the relative voltage drop across the contacts. The circuit proposed is arranged in such a way that both the input and output shafts are grounded, which allows the use of a single signal collection point. By using this method all of the critical contact capacitances are arranged in a parallel circuit, which in capacitive terms means they can be simply added together. Simulated results indicated that even if a single contact fails, the resulting effect on the total capacitance is significant enough to be immediately obvious. Even in terms of the measured relative voltages (V2/V1), the effect of single contact failure is easily noticeable.

CHAPTER 9: DESIGN CRITIQUE AND PROTOTYPE DESIGN

9.1 Design Critique

The purpose of a design critique is to determine the current stage and direction of the design process, whilst systematically assessing any potential problems and discussing how these can be resolved. One method of assessing a design is by comparing it to an "Ideal Final Result", which describes the solution to a technical problem, independent of the mechanism or constraints of the original problem (Domb, 1997). Hence the ideal final result of a CP-CVT would have an infinite ratio range, and zero relative mass. By comparing a design solution to an ideal final result, any design concerns that still need to be improved become immediately apparent. A comparison of the ideal final result to the current CP-CVT design is shown in Table 29.

Table 29: Comparison of Ideal Final Result and current design

Parameter	*Ideal Final Result*	*Current Design*	*Conclusion*
Mass	Negligible	As low as possible with current materials	Acceptable
Efficiency	Zero-Loss	>90%	Can be improved
Size	Negligible	As small as feasibly possible	Acceptable
Capacity	Unlimited	Capable of fulfilling specific applications	Acceptable
Ratio Range	Negative and infinite positive	Non-negative and limited positive	Can be improved
Ratio Control	Perfect control	Situation specific control	Can be improved
Durability	Indestructible	If design limits are imposed than durability is satisfactory	Acceptable if monitored

The early use of the Quality Function Deployment method and associated optimisation techniques improved aspects such as size, mass, and torque capacity as much as feasibly possible within the current design, and hence these design parameters are considered acceptable. Efficiency, ratio range and ratio-control ideally need to be improved however, which may require fundamental modifications or additions to the current design.

The problems found from a comparison with the ideal final design yielded the following design questions:

- Loading System: *Can a fixed loading system provide appropriate ratio control?*
- Geared Neutral: *How can the CP-CVT be adapted to increase ratio range and allow the vehicle to move from rest?*
- Customer Perceptions: *What can be done to pacify customer concerns about lack of control and perceived acceleration?*
- Cage Inefficiency: *Can the cage system be adapted to improve overall transmission efficiency?*
- In Situ Monitoring: *Can the system be monitored to ensure satisfactory lubrication and operation?*
- Traction Fluid Life: *Is the traction fluid expected to be changed regularly, and if so, how can the user determine how often this needs to happen?*

9.1.1 Loading System

It has been established that fixed loading system is appropriate for certain driving situations, but for safety and to return a level of control to the driver, this loading system needs to be adjustable in situ. In general there are three possible methods or providing an adjustable loading system:

- Hydraulic loading system
- Manual tightening of spring pre-compression
- Semi-automatic adjustment of spring pre-compression

A hydraulic loading system would remove the requirement of a spring completely and allow full control of the load applied to the toroidal disc, and hence in essence complete control of the transmission ratio at any point. Whilst this would potentially improve the vehicular performance and fuel efficiency by allowing the selection of the best transmission ratio at any moment, it also would increase the cost, complexity and losses within the CP-CVT. An alternative solution is to allow manual control of the spring pre-compression, which could, in its simplest form, consist of tightening a ‘loading nut’ on the shaft in which the spring is placed.

This would be more suited for applications where a continuous adjustment of the loading system is not required. A compromise would be to use a rotational actuator to control this 'loading nut', which could be adjusted either by the driver, or by an advanced algorithm depending on driving conditions. This maybe considered the best solution, since it would not be as expensive or complicated to implement in comparison to a hydraulic loading system, and assuming continuous adjustment is not required, would not detrimentally affect the transmission efficiency. However the implementation of this solution would possibly require a fundamental change the arrangement of the CP-CVT, as shown in Figure 9-1.

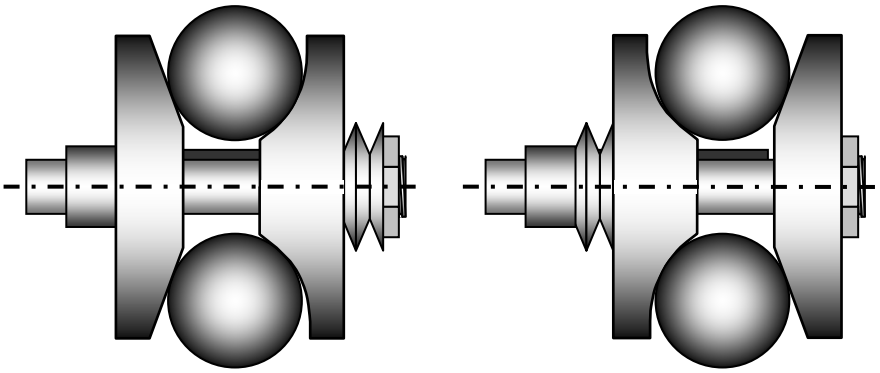


Figure 9-1: Alternative CP-CVT component layout

The original component layout that has been considered thus far assumes that the loading system is applied to the toroidal disc, which is situated towards the end of the input shaft and hence is more difficult to access. For a spring type loading system this is not an issue, however if additional loading control is required then it would be far more logical to swap the position of the conical and toroidal discs. This layout was considered early in the initial concept of the CP-CVT and abandoned due to the inferior torque capacity, increased spin losses and increased difficulty in fitting a cage (the channel of which must run parallel to the conical input disc surface). A simpler solution might be to apply the loading system to the conical input disc instead; however this drastically alters the torque response of the CP-CVT, which has already been determined to be effective in most situations. A hydraulic loading system would almost certainly require the alternative layout shown; however it is possible to fit an electronically-controlled actuator within the original layout without interfering with the cage system.

9.1.2 Geared Neutral

One of the requirements of an automotive vehicle's drive train is its ability to disconnect the engine from the wheels, which is required when the vehicle is at rest. Despite this, previous literature regarding the CP-CVT has not substantially addressed this requirement. Other traction drives often employ a planetary/epicyclic gear system within their design allowing 'geared neutral' and reverse motion, whilst the vehicular simulation described assumes the use of an automated-mechanical clutch, similar to those used in dual-clutch transmissions. Both of these solutions are reasonable, and each has its advantages. Automated clutches increase the overall weight of the transmission system, but add an additional point of control. In comparison to torque convertors they are very efficient; hence the transmission efficiency and fundamental behaviour of the CP-CVT would be unaffected. Planetary gear systems conversely alter a number of properties of the CP-CVT, including the transmission ratio response, ratio range, and torque capacity. Despite this, planetary gear systems are generally favourable since they extend the range of ratio's available, producing what is normally known as an infinitely variable transmission (IVT). The concept of using a planetary gear-system within the CP-CVT has been previously analysed, as shown in Figure 6-16, which would allow both idling and reverse. Additionally, a number of design concepts were considered, one of these, which also incorporates a double cage concept is shown in Figure 9-2.

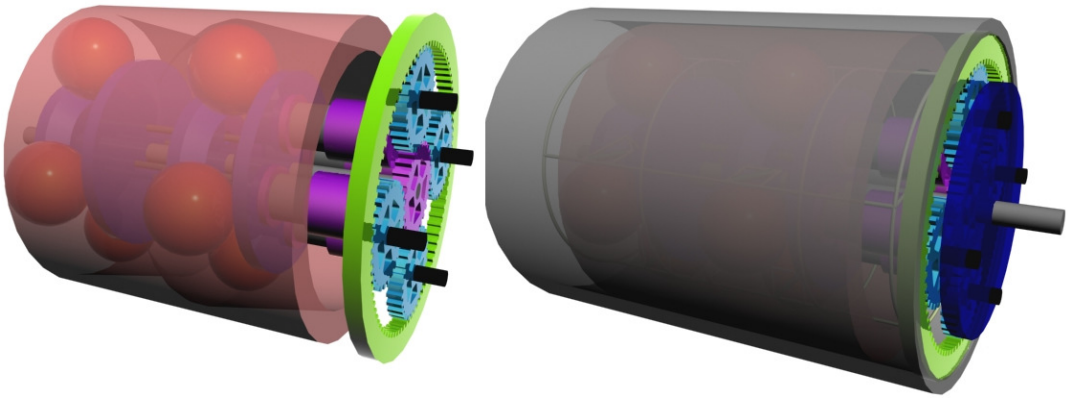


Figure 9-2: Early concept idea for planetary gear system

The concept of a planetary gear system was disregarded at this stage until the precise behaviour of the CP-CVT can be determined. A planetary gear system would split the resistive torque applied between the CP-CVT output and engine, making the behaviour of the CP-CVT depend entirely on the specifics of the particular engine, making simulation almost impossible. Furthermore, to allow reverse motion the toroidal disc position would have to be precisely

controlled, which is not currently possible or desirable. The use of a planetary gear-system will be reconsidered at a later stage once prototype testing has been conducted.

9.1.3 Customer Perceptions

Customer perception of CVT technology remains largely negative, partly due to early design flaws, and partly because of the difference in driving 'feel' that occurs in vehicles fitted with a CVT. When accelerating, drivers generally expect to hear a corresponding increase in engine noise, which partly aids the perception of acceleration. Unless a CVT is designed to mimic this increased engine noise, a driver usually feels that the acceleration of the vehicle is inferior, even when this is untrue. This is such a concern to CVT designers that certain ratio control strategies take this into account, as discussed in the literature review. A similar concern is that if a CVT is designed to allow the engine to operate as efficiently as possible then performance can suffer, and hence a compromise must be made. Alternatively the ratio-control strategy can be adaptable depending on throttle position, etc, to provide either higher engine efficiency or high engine power output. Any complex ratio-control such as this would invariable requires the use of a hydraulic-type loading system, which as discussed earlier, is not an attractive solution within the CP-CVT.

The vehicular simulation of the CP-CVT indicated that it automatically adjusts the transmission ratio based on the engine torque output and hence throttle position at any moment. Hence in reality the CP-CVT already incorporates different operating 'modes' depending on throttle position, providing either higher power output when the torque output is high, or efficient engine operation when torque increase is applied steadily. Whether this would be reflected and indeed acceptable in reality would require a prototype CP-CVT vehicle and extensive driver perception testing. In any case it is clear that if a driver wishes to improve fuel economy then they will have to adapt to a change in transmission technology and hence driving style.

9.1.4 Cage Inefficiency

It was established earlier that the interface between the ball elements and the separator/cage accounts for the largest proportion of the losses within the CP-CVT. Even though the overall transmission efficiency has been deemed to be acceptable, this still remains a concern. High energy losses at this interface will be directly converted into heat energy, increasing the lubricant temperature. This could potentially decrease film thickness, increasing the likelihood of wear or even failure. Assuming that it is not desirable to use an additional lubricant cooling system, there remains a limited number of solutions to this problem.

The simplest type of cage design is shown in Figure 9-3, which simply consists of a concavely curved surface. This curved surface serves to increase the contact area between the separator and the ball element, decreasing contact pressure and hence losses.

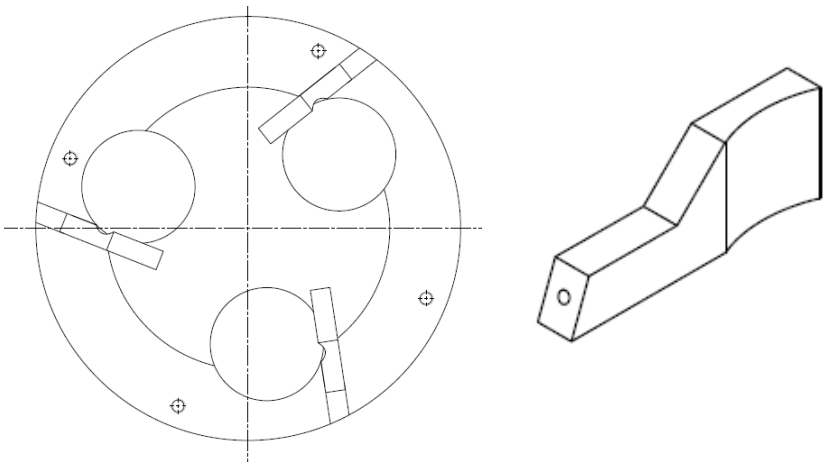


Figure 9-3: Example of simple ball separator design

More complex designs for this separator were also considered which allow an axis of rotation of the contacting surface, reducing the losses further; however the space required for these types of separators significantly limit the component dimensions, and furthermore add an additional point of failure, making their use unattractive. One possible method of solving this problem is shown in Figure 9-4. This revolutionary design, which was developed and patented by the original CP-CVT designer, replaces the ball elements with two hemispheres, attached to a movable shaft, eliminating the need for a cage altogether.

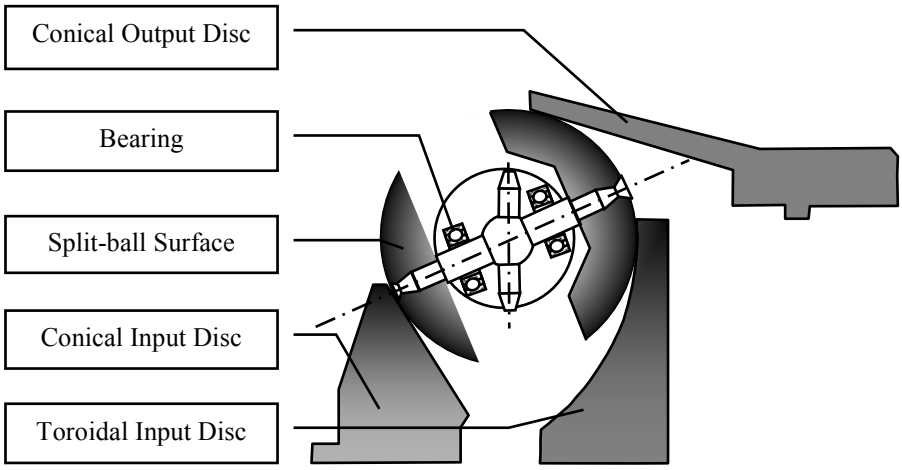


Figure 9-4: Complex ball design to eliminate cage requirement

The complexity of this design and the move away from a fundamentally strong spherical structure means this concept is unsuitable for any high torque applications, such as automotive vehicles.

The best solution hence remains the simple separator shown in Figure 9-3, both in terms of cost, simplicity and reliability. It is entirely plausible that the losses at this interface can be reduced through optimisation of the surface geometry and the use of a low-friction coating; however the magnitude of the savings that can be made in terms of loss cannot be easily studied without a prototype test-rig.

9.1.5 In Situ Monitoring/Traction Fluid Life

The proposed capacitive measurement system described in Chapter 8 is one possible method of monitoring the film thickness within the CP-CVT during operation. This would allow the film thickness to be monitored at all times, ensuring that it remains within a predetermined limit. This would be especially useful during start up as the contact develops, and during extended vehicle use where it is possible that the lubricant temperature would increase unacceptably resulting in a significantly reduced film thickness. Typically this is not considered by traction drive designers whom generally assume that traction fluid behaviour can be predicted based on laboratory tests, including tests that run until failure. These tests have provided useful information regarding operating limits and it is now generally understood how often the lubricant will need to be changed, and how long the critical components are expected to last. However until traction drives are widely implemented these results are based only on a relatively small number of isolated tests, which do not always simulate driving conditions, hence there are potentially other reasons for lubricant failure that have not yet been considered. This implies that a film thickness monitoring system (such as the one proposed) may ultimately be invaluable in ensuring traction drives such as the CP-CVT have reasonable operating life, or conversely, if laboratory test results are found to be repeated during vehicular operation, they may never be required.

9.2 Bench-top Prototype Design

9.2.1 Component Dimensions

In order to test and validate the optimisation calculations and operational simulation, a scaled-down prototype test-rig was designed, although due to financial limitations it has not yet been constructed and hence will be the subject of future work. The test rig was designed to be used with a 0.37kW RS industrial DC motor and associated controller. The motor, when coupled with a 10:1 ratio gearbox has a maximum torque output of 11.8Nm at 300rpm. The dimensions of the CP-CVT were hence optimised for the reduced torque requirements using the optimisation techniques described previously. Since this is a one-off prototype, rather than a mass-produced product, the dimensions were adjusted and rounded for simplicity in manufacture. The resulting dimensions of the key components and algorithm-obtained technical properties are shown in Table 30 and Table 31 respectively.

Table 30: Prototype component dimensions

Parameter	*Value*
β	20°
γ	60°
R	0.02m
R_1	0.03m
r_0	0.04m

Table 31: Prototype technical properties

Technical Characteristic	*Value*
Overall Efficiency	92.9%
Normalised Transmission Ratio	4.21
Maximum Torque In	12Nm
Indicative Mass	2.26kg
Indicative Length	0.059m
Indicative Diameter	0.060m
Power Variation Coefficient	0.96

9.2.2 Ball-Screw and Output

The dynamic force calculations indicated that the ball screw lead-length would have to be less than 27.6mm to ensure that the normal forces would be sufficiently large that the traction coefficients at each contact would not exceed their design value of 0.045. As expected, this lead length is relatively large since the torque applied to the input is not particularly high, and hence only a small tractive force is required. Typically a ball screw's diameter increases linearly with the lead length, hence it is generally preferable to have a shorter lead length, which increases the force produced by the ball screw reducing the traction coefficient further. However any force produced by the ball screw has to be balanced by the force applied to the toroidal disc and hence the overall loads applied to each component increase as well. For a relatively small prototype this is a concern due to the limited material strength. Based on this, an SKF ball screw with a lead length of 20mm was chosen, which has a nominal screw diameter of 53mm.

The shaft of the ball screw-coupling, which forms the output shaft of the CP-CVT, is directly connected to the torque-loading system. This resistive torque is created by an eddy-current brake, which enables a controllable load through the use of a changing induced current. On this shaft there is also a torque transducer to monitor the rotational speed and torque at the output shaft, as shown in Figure 9-5.

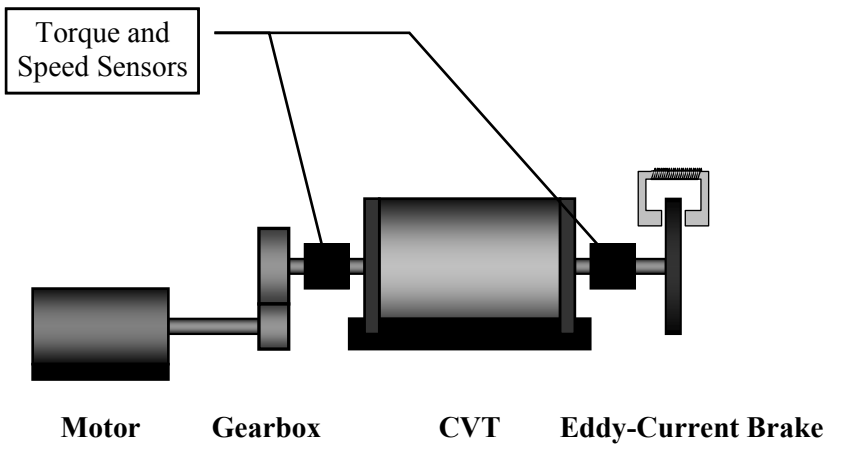


Figure 9-5: General arrangement of CP-CVT bench-top test rig

Rather than attaching the screw part of the ball-screw coupling directly to the conical output disc, it is instead attached to an output disc coupling as shown in Figure 9-6. This essentially allows the conical output disc to be manufactured in two separate parts: the contacting surface, which must be manufactured of expensive, bearing-quality steel, and the coupling, which can be manufacturer from a cheaper and lighter aluminium alloy.

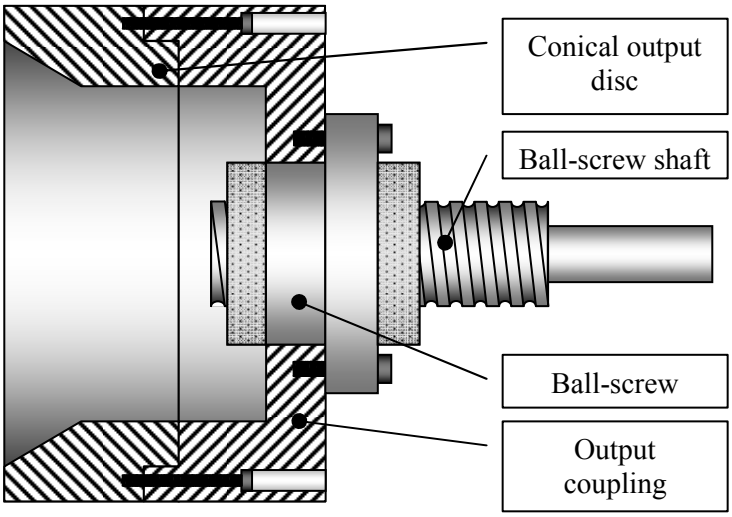


Figure 9-6: Conical output disc coupling

9.2.3 Input Discs

One side of the reduction gearbox is attached directly to the motor, whilst the other side contains an additional torque transducer to measure the torque and speed at the input side of the CP-CVT. The shaft containing the input discs has a stepped diameter, which prevents axial movement of the conical input disc. The toroidal input disc is loaded by a disc-spring, which can be pre-compressed via a simple nut attached to the end of the shaft, as shown in Figure 9-7. A keyway mechanism is used to ensure the discs rotate with the shaft.

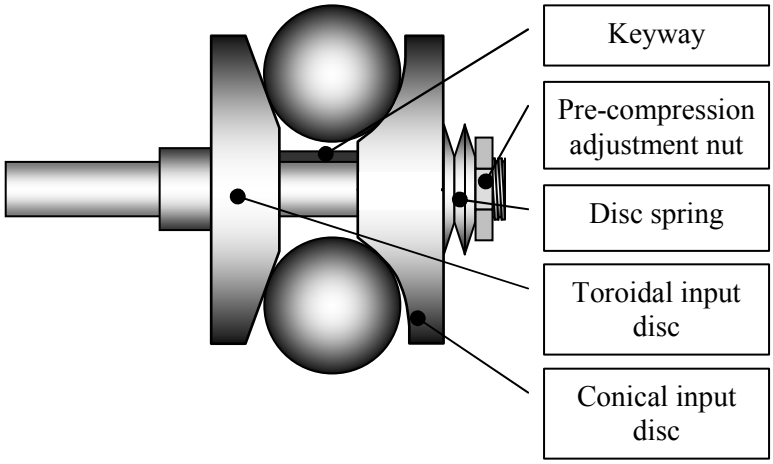


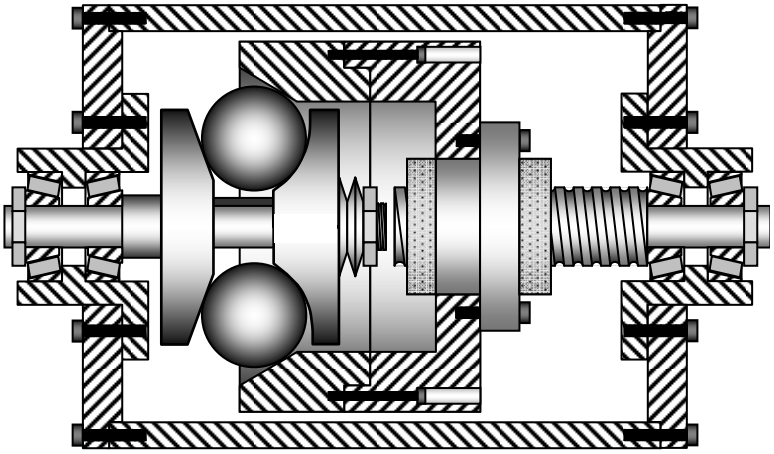
Figure 9-7: Input discs assembly

9.2.4 Casing

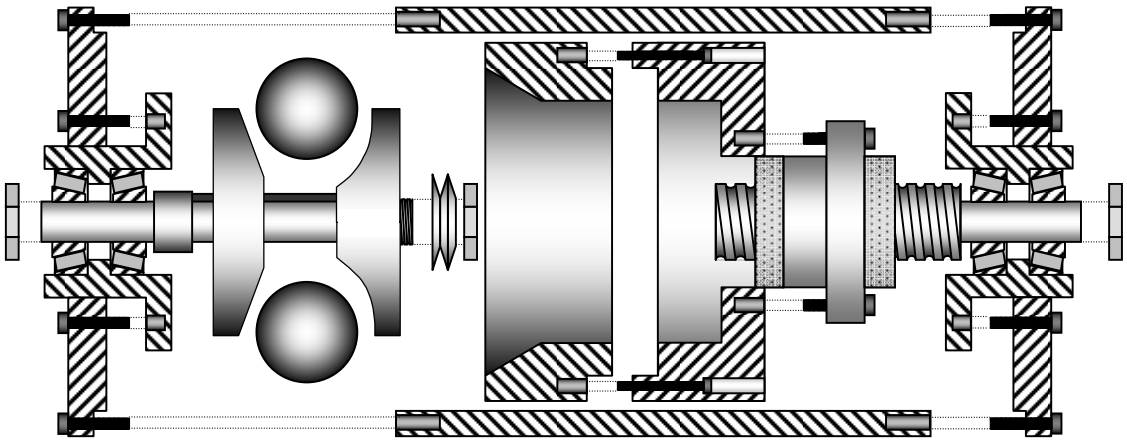
For simplicity the main casing consists of a cylindrical tube, with flat plates attached at each end. The end discs are identical, thus insuring alignment between the input and output shafts. At the centre of the end-plates are flanges that contain a pair of opposing angular-contact bearings allowing rotational freedom of the shafts but preventing axial movement. Gaskets are placed between the plates and the cylindrical tube, and between the flanges to prevent oil leakage.

9.2.5 Complete Assembly

A complete assembly of the proposed CP-CVT test rig, excluding the ball separator is shown in Figure 9-8, whilst an isometric engineering layout of the complete design is shown in Figure 9-9.



Assembled



Exploded

Figure 9-8: Assembly and exploded views of proposed prototype

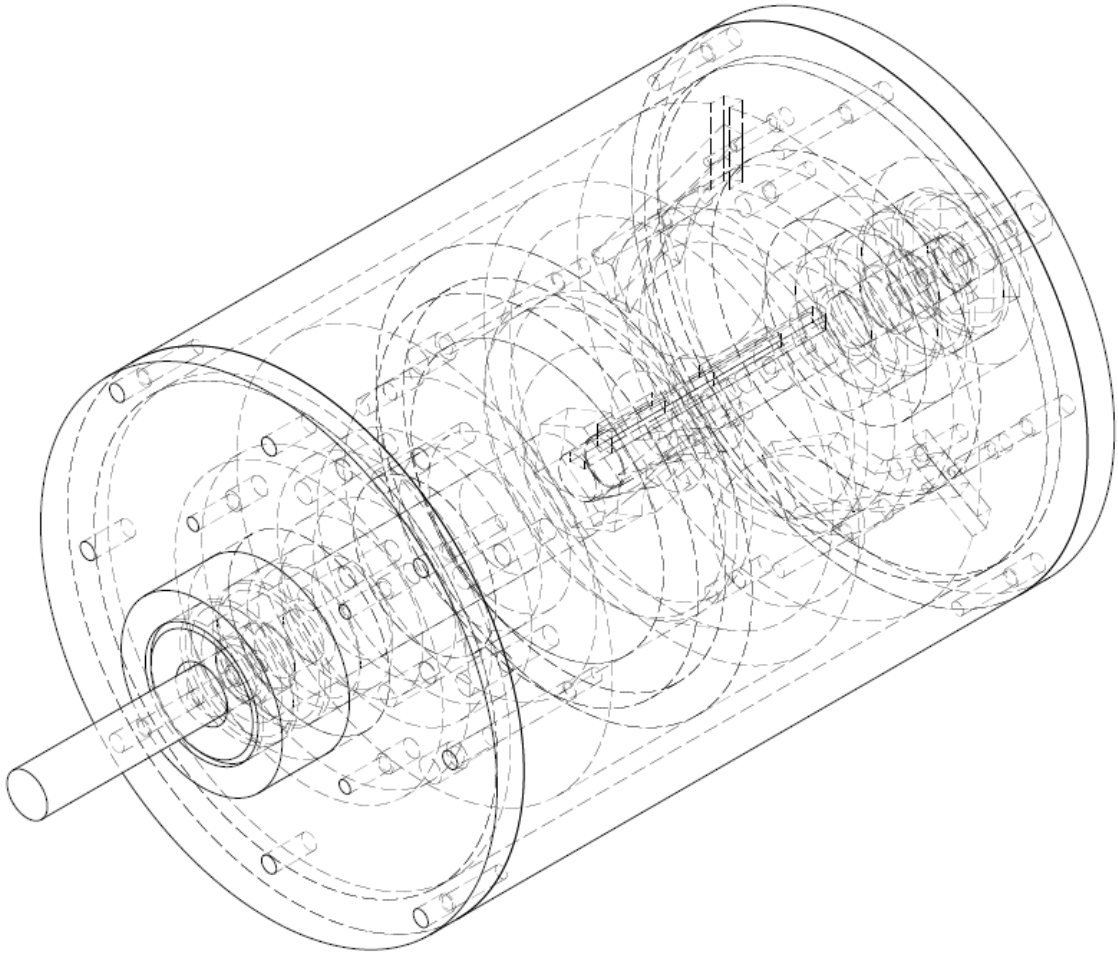


Figure 9-9: Engineering drawing of complete prototype assembly

Based on volumetric calculations of all of the parts, the total mass (assuming bearing steel is used for the ball elements, contacting discs and separator, and a lighter aluminium alloy is used for other parts except the shaft) was found to be approximately 10.5kg. This estimated mass excludes the traction fluid, bearings and ball screw. When these additional parts are included, the total weight of the CP-CVT prototype is approximately 19kg. This is a substantial increase over the indicative mass calculated previously (2.26kg), however as stated previously, this value was representative rather than absolute, consisting only of the mass of the contacting discs and ball elements. An increase in indicative mass leads to a proportionally similar increase in the total mass. Likewise, the overall diameter of the prototype is 170mm, whilst the length is 230mm, again an increase on the indicative dimensions found previously.

CHAPTER 10: CONCLUSION

10.1 Summary of Research Achievements

- The Quality Function Deployment process, which is not usually applied to transmissions systems, was successfully implemented to determine specific weighted targets for the technical requirements of the CP-CVT, based on their perceived impact on customer satisfaction.

- A novel optimisation algorithm was developed to determine the maximum theoretical efficiency of the CP-CVT design based on a number of formulae derived to assess the total losses within the CP-CVT or similar traction drive designs.

- Both the novel optimisation algorithm and a modified genetic algorithm were adapted to dimensionally optimise the CP-CVT components based on a number of technical requirements. By incorporating the QFD process within the algorithm targets a complete optimisation process was developed to directly determine the influence of design choices and component dimensions on overall customer satisfaction.

- A unique vehicular simulation tool was created from fundamental theories and models that was capable of accurately representing real vehicle performance when coupled with a standard step gear-transmission. By incorporating the CP-CVT within the simulation, certain aspects of the design (such as the loading system) were optimised and assessed without the need for costly empirical testing.

- By using a specially designed tire friction test rig, it was found that rubber friction can depend more on the contact pressure than the contact area, implying that friction is governed more by the material hysteresis than surface adhesion. It was also found that the magnitude of the improvement in friction force that can be obtained by careful selection of macroscopic roughness is comparable to using different microscopically rough surfaces.

- An innovative in-situ monitoring system was proposed and designed based on the theory of parallel plate capacitance. The use of this system in the CP-CVT or similar traction drives could potentially give an advance warning of lubricant failure, reducing wear and extending the life of the transmission system.

- It was ultimately determined that through the design tools employed, the CP-CVT is now ready for prototype development. A specific prototype test-rig was consequently designed using the optimisation process developed earlier, and is ready for construction.

- Although the optimisation process was applied to the CP-CVT, the design tools and methods developed can be easily adapted and applied to a wide variety of transmission designs, helping to optimise them both in terms of customer satisfaction and dynamic performance.

10.2 Overview of Thesis

10.2.1 Quality Function Deployment as a Design Tool

After the initial literature review, it was determined that there is a definite market need for an improved continuously variable transmission design. The need for CVT technology within the automotive industry arises from the variable operating efficiency of internal-combustion engines. In order to improve overall fuel efficiency, and hence reduce emissions, an engine requires a continuously variable transmission ratio to allow it to remain in a peak-efficiency operating envelope for prolonged periods of time. Current CVT designs are fairly well developed, but generally suffer from flaws such as complex control mechanisms, or limited torque capacities that have hindered their penetration and acceptance within the automotive industry.

A unique traction-type CVT design has previously been designed (CP-CVT), which overcomes several of these flaws, making it potentially a very strong candidate in the race for the next widely-implemented automotive transmission design. At the start of this research, the CP-CVT design was still in the early stages of development and no previous attempt had been made to improve it with respect to customer and technical demands, which is required if it is to succeed.

To improve the CP-CVT design, initially a process known as Quality Function Deployment was employed, with the objective of improving the design in terms of specific, measurable technical requirements. These technical requirements were derived and related to fundamental customer demands. This approach is thus a form of customer-orientated design, which is normally applied to parts and products in which the customer has a direct interaction, and hence has not been extensively applied to a CVT-type transmission system. Despite this, the method was successfully applied to the CP-CVT, yielding technical design requirements that were prioritised according to their perceived influence on overall customer satisfaction for a variety of specific applications using a variation of the House of Quality technique. It was concluded that the most significant improvement in technical characteristics could be achieved through optimisation of the component dimensions.

10.2.2 Efficiency Optimisation

Dimensional optimisation of the CP-CVT initially focused only on transmission efficiency, which was found to be one of the most critical requirements, and also the most complex to calculate and predict. The transmission efficiency of the CP-CVT was shown to be a function of several different sources of loss, including bearings, churning, cage and the EHD contact. A series of equations were developed for each of these loss sources based on existing literature to

quickly predict the overall transmission efficiency of the CP-CVT for any particular torque and speed input, and any combination of component dimensions.

It was found through test calculations on existing component dimensions offered in previous literature that the current efficiency of the device is relatively poor. This was improved through the use of a novel search algorithm, which yielded efficiency values that are more comparable to alternative traction drive designs.

Although the efficiency could be considered representative, rather than absolute, the overall calculated efficiency of the CP-CVT is a significant improvement on the automatic transmissions currently used in the automotive industry and a slight improvement over alternative CVT designs. One of the reasons for this is the absence of side-slip within the contact and the lack of an actuator to control the transmission ratio and a hydraulic loading system to provide the normal force required, both of which decrease overall efficiency.

The novel search algorithm created (fuzzy-swarm), which is based on a combination of ant-colony and particle swarm techniques, performed very well, despite its relative simplicity and ease of implementation. One of the reasons for this is that the algorithm is quickly able to map the search space and determine stronger regions with only a limited number of evaluations, hence offering a 'fuzzy' rather than exact set of results. This makes it a very useful 'first-step' in single-criteria optimisation. Although similar in name, the fuzzy-swarm technique developed is distinctly different from existing optimisation methods, requiring no evaluation of adjacent solutions, significantly reducing computational time.

10.2.3 Multi-criteria Optimisation

Once the transmission efficiency was found to be acceptable, the next stage involved optimising all of the technical requirements simultaneously. To achieve this, a number of evaluation equations were derived for each of the technical requirements, based on the fundamental behaviour of the CP-CVT. A method was created to score each set of dimensions on their ability to meet a set of application-specific targets. The target values and relative weightings were derived for three specific automotive applications through the use of different aspects of the House of Quality and general quality function deployment techniques. The relative scores of each individual requirement were then combined using a Pareto-type function, yielding an overall fitness score. Since this score was based on the customer-orientated House of Quality technique, this overall score can be seen as a direct measure of the fulfilment of the original customer demands.

Initially the component dimensions were optimised using a multi-criteria version of the fuzzy-swarm algorithm developed previously, demonstrating the flexibly of this novel optimisation technique. Using this technique, several sets of dimensions were presented that fulfilled the customer demands to an acceptable level, however the variation of the results indicated that the search technique was not accurate enough. Hence to determine precise results, a more complicated and computationally more demanding optimisation algorithm was used, based around genetic evolution. This complex algorithm utilises the principle of natural selection to evolve and improve a set of results across several hundred evolutions, theoretically culminating in the strongest possible set of input dimensions. By using this algorithm the overall Pareto fitness score was improved by reducing certain technical characteristics that previously exceeded the targets (over-engineered) whilst improving other areas. By using this technique it was concluded that although overall efficiency is considered technically important, a greater overall customer satisfaction can be achieved by sacrificing this in favour of improving other technical aspects.

Neither optimisation algorithm was initially perfect, producing results that had some degree of variation and hence inconsistency when the algorithm was repeated using identical conditions. In order to increase the consistency, several improvements were made to each algorithm. The genetic algorithm was improved through the use of parallel evolution, which repeats the algorithm several times, seeding good solutions each time in the initial population. By reducing the number of evolutionary stages and population size, the overall computational time was found to be nearly identical to a standard genetic algorithm. Despite this, the modified genetic algorithm was still far more complex to implement and computationally more demanding in comparison to the fuzzy-swarm algorithm. Improvements were hence made to the fuzzy-swarm method to reduce the fuzziness of the results, whilst still maintaining its fundamental simplicity and relatively fast computational time. This was achieved by stochastically restarting the algorithm then seeding the subsequent algorithm with the strongest solutions found thus far. By doing this, the algorithm focused its search on an iteratively decreasing search region, culminating in solutions that were found to be near-identical to those produced by the genetic algorithm. Although this adaptation could conceivably lead to clustering around false-maxima, this was not found to occur with this particular optimisation problem.

By using these algorithms an optimised set of application-specific dimensions was produced. Although the dimensions yielded are specific to the applications discussed, the optimisation processes can be easily adapted to produce an optimised set of dimensions for any particular application, automotive or otherwise, simply by filling in a House of Quality.

10.2.4 Vehicular Simulation

Once a set of optimised dimensions was determined, the dynamic behaviour of the CP-CVT was assessed in a vehicular environment. This was achieved through the use of a specially designed simulation, which was created using fundamental equations, based on existing empirical and theoretical models, resulting in a fully-customisable, user-friendly simulation tool. Initially the simulation was run with an automatic transmission replacing the CP-CVT, which allowed a comparison of the simulated results with real performance figures. By using appropriate vehicle and engine parameters, which were obtained from manufacturer specifications, a good correlation was found between the simulation and real behaviour, validating several aspects of the model such as engine behaviour and tire friction.

The CP-CVT was then implemented within the simulation using component dimensions that had previously been optimised for use in a family car. An initial comparison of the vehicular performance between a standard step-gear transmission and the CP-CVT indicated that the stated performance improvements offered by a CVT were not attainable when the CP-CVT is entirely autonomously controlled. Despite this, the acceleration and top speed values were comparable, with the added benefit that vehicular acceleration is smoother and the engine generally operates at a lower speed and higher torque output, reducing noise and theoretically improving fuel efficiency.

By analysing the loading system of the CP-CVT it was found that the selection of an appropriate spring stiffness and pre-compression is perhaps even more critical than component dimensions in terms of ratio control. It was also found that the ratio of the CP-CVT is controlled by the engine torque applied, and hence the throttle position. In partial throttle conditions, the highest transmission ratio is not utilised, meaning the engine speed remains lower (generally more efficient), whilst in open-throttle conditions a higher ratio is used, increasing engine speed and torque output, providing improved acceleration albeit at the expense of fuel economy.

Through the use of the simulation, it was concluded that a fully torque controlled CVT is theoretically possible, however realistically an additional level of control would be required for safety and in specific situations such as slippery road conditions or when descending a steep gradient where the use of ‘engine braking’ is preferable. The constant power-aspect of the CP-CVT is still somewhat ambiguous, since its behaviour depends more on the specified loading system, which can be as simple or complex as the application requires. Despite this, the power-variation coefficient did appear to have an influence on the CP-CVT’s ability to provide an ideal transmission ratio since it was able to utilise a wider range of ratios.

Dimensional optimisation established that the CP-CVT is capable of fulfilling specific technical requirements, whilst the simulation has shown that the CP-CVT can also perform very well dynamically.

10.2.5 Complementary Research

In addition to the core design process, additional complementary research was also conducted in the associated areas of rubber tire friction and film thickness measurement.

During the creation of the vehicular simulation it was found that traditional tire friction models (even those that claim to be based upon physical phenomena) are invariably only useful in simulation through 'trial and error' data fitting. Whilst they can be used to estimate the effect of load and velocity on friction, they cannot be used to predict the effect of physical phenomenon, and furthermore cannot be used to theoretically improve tire friction. Complementary research thus assessed the validity of existing tire friction models in representing physical phenomenon such as macroscopic roughness. By using a specially design tire-friction test rig it was found that macroscopic asperity radius and pitch length have a large influence on the deformation friction and hence the overall friction coefficient. It was determined that a higher friction force could be achieved at lower slip ratios through the use of smaller radius asperities, whilst at higher slip ratios the distance between asperities has a stronger effect. It was concluded that friction depends more on the contact pressure than the contact area, implying that the friction is governed more by the hysteresis (deformation) term than the adhesive term under the contact conditions tested. Hence by approximating the contact pressure using Hertzian theory, a very reasonable prediction of the overall effect of altering either the asperity radius or pitch length can be found. It was also found that the magnitude of the improvement in rubber friction that can be obtained by careful selection of macroscopic roughness is comparable to using different microscopically rough surfaces.

The other complementary research chapter proposed a method for the measurement and monitoring of the film thickness of the EHD contacts within the CP-CVT during operation. The proposed method utilises the theory of parallel-plate capacitance, which states that the distance separating the plates (the film thickness in this case) is inversely proportional to the capacitance measured. Hence if the film thickness decreases due to abnormal temperature rise or lubricant starvation, the capacitance across the contact would increase substantially. The proposed method used Hertzian theory to determine the contact area, whilst the permittivity of the lubricating oil was determined by a series of capacitance tests conducted on a modified PCS Instruments test rig. The test rig measured the capacitance of an EHD contact formed between a steel disc and a steel ball, which was then compared to film thickness results found using a

known optical technique in identical EHD conditions. Through these tests, it was concluded that the permittivity of the oil increases by a factor of approximately 3 due to the increased contact pressure. The proposed monitoring system ensures all of the critical contact capacitances are arranged in a parallel circuit. By using this arrangement, simulated results indicated that even if a single contact fails, the resulting effect on the total capacitance is significant enough to be immediately obvious, thus providing a potentially invaluable tool for monitoring the film thicknesses within the CP-CVT during operation.

10.2.6 Design Critique and Prototype Design

Through extensive design optimisation and investigation, the CP-CVT was proven to possess all the necessary technical requirements for automotive purposes. A design critique performed on the CP-CVT showed that a number of initial design problems had been addressed and improved through the use of the optimisation and design techniques discussed. The critique also indicated that a number of problems still need to be addressed for which several solutions have been offered and discussed. It was ultimately concluded that the CP-CVT is ready for prototype development. A specific prototype test-rig was consequently designed using the optimisation processes developed earlier, and is ready for construction. If the proposed refinements are included and the prototype test results verify the conclusions found, then the CP-CVT is ready for testing in an automotive vehicle environment.

10.3 Suggestions for Future Work

10.3.1 Short-Term

- Further assessment of the loading system is required to determine if additional loading control is required, and how it can be adapted for use in multiple situations.
- The simulation should ideally be enhanced to incorporate additional driving situations (such as gradients) and allow real-time simulation, ideally in a driving simulator, thus providing the potential for an assessment of perceived drivability.
- Additional development of the fuzzy-swarm optimisation algorithm could prove to be exceptionally beneficial to a number of different industries.
- The proposed capacitance monitoring system to traction-drive designs needs to be experimentally validated, with the ultimate aim of becoming easily implemented in a any application that utilised standard traction lubrication.

10.3.2 Long-Term

- The bench-top prototype that has been designed needs to be constructed to determine the validity of the technical calculations and efficiency predictions. Furthermore, by using a controllable eddy current brake a comparison can be made between the simulated vehicle performance and the behaviour of the prototype.
- Ultimately a vehicle-implemented prototype needs to be constructed using the optimisation processes developed to verify the conclusions shown, and also to assess and improve the drivability performance of the CP-CVT before production can commence.
- Additional applications of the CP-CVT need to be researched in order to improve its potential for further investment and development.
- The tools and methods used through this research would ideally be woven together into a single, universal optimisation tool that could be applied to a number of design problems, in addition to automotive transmission systems.

VIII. REFERENCES

Akehurst, S., Brace, C.J., Vaughan, N.D., Milner, P. and Hosoi, Y., 2001. "Performance Investigations of a Novel Rolling Traction CVT", *SAE 2001 World Congress*, USA, March 2001, Society of Automotive Engineers, Inc.

Akehurst, S., Parker, D.A. and Schaaf, S., 2007. "Dynamic Modeling of the Milner Continuously Variable Transmission - The Basic Kinematics", *Journal of Mechanical Design*, Vol. 129, Iss. 4, pp.1170-1178.

Akehurst, S., Moyers, J., Hunt, A. and Schaaf, S., 2009. "Development of a Design Tool for Modelling and Optimisation of the Milner CVT", *Proc. Of MPT2009, JSME International Conference on Motion Power Transmissions*, May 2009, Sendai, Japan.

Akao, Y., 1990. *Quality function deployment*, Productivity Press, Cambridge, UK.

Andrzej, O. and Stanislaw, K., 2000. "A New Constraint Tournament Selection Method for Multi-criteria Optimization Using Genetic Algorithm", *The 2000 Congress on Evolutionary Computation CEC 00*, California, USA, July 2000, pp501-508

Anghel, V., Glovnea, R.P. and Spikes, H.A., 2004. "Friction and film-forming behaviour of five traction fluids", *Journal of Synthetic Lubrication*, Vol. 21, No. 1, pp13-32

Archer, J.F., 1957. "Elastic deformation and the laws of friction," *Proceedings of the Royal Society of London, Series A,* pp190-205.

Arita, M., 2000, "Recent CVT Technology and Their Effect on Improving Fuel Economy", *Proceedings International Tribology Conference*, Nagasaki, Japanese Society of Tribologists, pp.197–201.

Askwith, T.C., Cameron, A. and Crouch, R.F. 1966. "Chain Length of Additives in Relation to Lubricants in Thin Film and Boundary Lubrication", *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, Vol.291, No.1427. pp.500-519.

Bai, H. and Kwong, C.K., 2003. "Inexact genetic algorithm approach to target values setting of engineering requirements in QFD", *Int. J. Prod. Res.* Vol.41. No.16. pp.3861-3881

Bakker, E., Nyborg, L. and Pacejka, H.B., 1987. "Tyre Modelling for Use in Vehicle Dynamics Studies" *SAE Technical Paper*, SAE Paper No. 870421.

Baudoin, P., 1979, "Continuously Variably Transmissions for Cars with High Ratio Coverage" *Continuously Variable Transmissions for Passenger Cars*. SAE. Warrendale, USA: Society of Automotive Engineers, Inc.

Bell, C., Glovnea, R.P. and Mares, C., 2008, "Concept Design for Transmission Systems", *Proceedings of International Multi-Conference on Engineering and Technological Innovation: IMETI*, June-July 2008, Orlando, FL, USA.

Bell, C. and Glovnea, R.P., 2009. "Modelling and Simulation of a Novel Toroidal-Type CVT", *Proc. Of MPT2009, JSME International Conference on Motion Power Transmissions*, May 2009, Sendai, Japan.

Bell, C.A., and Glovnea, R.P., 2011. "Tribological Optimisation of a Toroidal-Type CVT", *Proceedings of the IMechE, Part J, Journal of Engineering Tribology*. (accepted for publication)

Bell, C.A., Mares, C. and Glovnea, R.P., n.d. "Concept design optimisation for Continuously Variable Transmissions", *Int. J. Mechatronics and Manufacturing Systems*, Vol. 4. No. 1. (publication confirmed).

Bharat Book Bureau, 2006. *Nearly 80% Of Passenger Cars Are Sold With Manual Transmission In Europe*, Bharat Book Bureau: Mumbai, India.

Biles, J.A., 1994. "GenJam: A genetic algorithm for generating jazz solos", *Proc. of Int. Computer Music Conf. (ICMC'94),* Aarhus, Denmark, 1994. pp. 131–137.

Boness, R. J., 1989. "Churning losses of disks and gears running partially submerged in oil", *Proceedings of 1989 International Power Transmissions and Gearing Conference*, Chicago, Illinois, Vol. 1, pp.255-359.

Boos, M., and Mozer, H., 1997. "Ecotronic: The Continuously Variable ZT Transmission (CVT)", *Transmission and Driveline Systems Symposium*. SAE Paper No. 970685, in SP-1241. pp.61-67.

Brockbank, C., 2009. "Fuel consumption improvements by applying IVT technology to commercial vehicles", *Advanced Transmissions for Low CO2 Vehicles*, Rueil Malmaison, France. June, 2009.

Brockbank, C. and Heumann, H., 1999. "Delivery of IVT for a 5 litre SUV: Addressing the Concerns of Geared Neutral", *IMechE Autotech Conference, UK*. November 1999.

Brockbank, C. and Greenwood, C., 2009. "Fuel economy benefits of a flywheel and CVT-based mechanical hybrid for city bus and commercial vehicle applications", *SAE Commercial Vehicle Engineering Congress*, Rosemount, USA. October, 2009.

Bucknell, J., 2006. "Powertrain Matching" *SAE Student Lecture*, [Online]. Available: www.sae.org/students/presentations/powertrainm.ppt [Last Accessed 06/08/2010].

Burtt, D.J., 2007. "Fuel economy benefits of a high torque, infinitely variable transmission for commercial vehicles", *SAE 2007 Commercial Vehicle Engineering Congress and Exhibition*, Rosemount, USA, October, 2007.

Burtt, D. and Hough, M., 2007. "Torotrak toroidal traction drive continuously variable transmission (T-CVT) for Kei cars", *CVT-Hybrid 2007 International Congress on Continuously Variable and Hybrid Transmissions*, Yokohama, Japan. September, 2007.

Cahn-Speyer, P., 1957. "Mechanical infinitely variable speed drives", *The Engineer's Digest*, Vol. 18, No.2, pp.41-65.

Cameron, A., 1985. "Righting a 40-year-old wrong: A. M. Ertel — the true author of 'Grubin EHL' solution" *Tribology International*, Vol. 18, Iss. 2, pp92.

Canudas-de-Wit, C., Tsiotras, P., Velenis, E., Basset, M. and Gissinger, G., 2002. "Dynamic Friction Models for Road-Tire Longitudinal Interaction" *Vehicle System Dynamics*.

Carson, R.W., 1975, "Focus on traction drives: 100 years of traction drives", *Power Transmission Design*, Vol. 17. No. 9. pp.84-88.

Caruana, R.A., Eshelman, L.J. and Schaffer, J.D., 1989. "Representation and hidden bias II: Eliminating defining length bias in genetic search via shuffle crossover". *Proceedings of the Eleventh International Joint Conference on Articial Intelligence*, pp.750-755.

Childs, P., 2004. *Mechanical Design (2nd Edition)*. Elsevier Ltd: Oxford, UK.

Chittenden, R.J., Dowson, D., Dunn, J.F. and Taylor, C.M., 1985. "A theoretical analysis of the isothermal elastohydrodynamic lubrication of concentrated contacts" *Proc. Roy. Soc. Lond. A397*, pp.245-269

Cretu, O.S., and Glovnea, R.P., 2003. "Traction Drive with Reduced Spin Losses", *ASME Trans. J. Trib*., Vol. 125, No.3, pp.507-512

Cretu, O.S., and Glovnea, R., 2005. "Constant Power Continuously Variable Tranmission (CP-CVT) Operating Principle and Analysis", *Journal of Mechanical Design*, Vol. 127, No.1, pp.114-119.

Cretu, O.S. and Glovnea, R.P., 2006. "Geometrical and Dimensional Optimisation of a Constant-Power Continuously Variable Transmission", *Buletinul Intsitutului Politehnic Iasi (Bulletin of the Technical University Iasi),* Vol. LII, Part 6A, pp.149-161

Dahl, P.R., 1976. "Solid Friction Damping of Mechanical Vibrations", *AIAA Journal*, Vol.14, No.12. pp1675-1683.

DCTfacts.com, 2008. *Why dual clutch technology will be big business*, [Online Report]. *dctfacts.com report archive*. Available from: http://www.dctfacts.com/archive/2008/why-dual-clutch-technology-big-business.aspx

De Jong, K., 1975. "An Analysis of the Behaviour of a Class of Genetic Adaptive Systems". PhD Thesis, University of Michigan, Ann Arbor, MI, USA. [Online]. Available: http://cs.gmu.edu/~eclab/kdj_thesis.html [Last Accessed 06/08/2010].

Denny, D.F., 1953. "The Influence of Load and Surface Roughness on the Friction of Rubber-Like Materials". *Proc Phys. Soc. LXVI*, pp.721-727

Deur, J., 2001. "Modeling and Analysis of Longitudinal Tire Dynamics Based on the LuGre Friction Model", *Proceedings of the 3rd IFAC Workshop on advances in Automotive Control, Vol. 1,* March 2001. pp101-106.

Dick, E., 2010. "The role of variable drive technology in realising fuel economy and emissions improvements", *FISITA World Automotive Congress*, Budapest, Hungary, May, 2010.

Dixon, J., 1988. "On research methodology towards a scientific theory of engineering design", *Design Theory '88*, Newsome S. L., Spillers W. L. and Finger S. (Eds.), Springer-Verlag.

Domb, E., 1997. "The Ideal Final Result: Tutorial". *The TRIZ Journal*, February 1997, 02-a.

Dorigo, M., 1992. *Optimization, Learning and Natural Algorithms*, PhD thesis, Politecnico di Milano, Italy.

Dorigo, M., Birattari, M. and Stutzle, T., 2006. "Ant Colony Optimization", *Computational Inelligence Magazine, IEEE*. Vol 1. Iss. 4. pp.28-39.

Dutta-Roy, T., 2004. *Effect of Continuously Variable Unit on Powertrain Dynamics*. (Thesis), The University of Technology, Sydney, Australia.

Evans, C.R. and Johnson, K.L., 1986, "The Rheological Properties of Elastohydrodynamic Lubricants", *Proc. I.Mech.Eng.,* Vol. 200, C5, pp.303-312.

Evans, S.P. and Lee, A., 2007, "Development of traction fluids for a range of applications", *CVT-Hybrid 2007 International Congress on Continuously Variable and Hybrid Transmissions*, Yokohama, Japan. September, 2007.

Evans, S., Lee, A., Hillsden, A. and Nagatomi, E., 2009. "The durability of traction fluid in full-toroidal traction drives under extreme high temperature conditions", *World Tribology Congress*, Kyoto, Japan, September 6-11, 2009.

Field, M. and Burke, M., 2005. "Powertrain control of the Torotrak infinitely variable transmission", *SAE 2005 World Congress & Exhibition*, Detroit, USA. April, 2005.

Fonseca, C.M. and Fleming, P.J., 1995. "An Overview of Evolutionary Algorithms in Multiobjective Optimization", *Evolutionary Computation*, Vol 3. Iss.1, pp.1-16.

Fuchs, R. and Hasuda, Y., 2004. "Dynamic Performance Analysis of a Full Toroidal IVT - A Theoretical Approach", *2004 International Continuously Variable and Hybrid Transmission Congress,* September 2004, University of California (UC Davis).

Fuchs, R., Hasuda, Y. and James, I., 2002. "Modeling, Simulation and Validation for the Control Development of a Full-Toroidal IVT" *Technical Report* (2002), Torotrak Development Ltd., Leyland, UK

Gim, G. and Nikravesh, P.E., 1999. "A Unified Semi-Empirical Tire Model with Higher Accuracy and Less Parameters," *SAE International Congress and Exposition,* 1999, Detroit, MI, USA.

Gilchrist, A., Earley, J.E. and Cole, R.H., 1957, "Effect of Pressure on Dielectric Properties and Volume of 1-Propanol and Glycerol", *Journal of Chemical Physics*, Vol. 26, Iss.1., pp196-200.

Gilson, D.R., 1998. "Using a Schering A.C. Bridge at low frequency to determine condenser properties", *Hull University Physics Teaching Laboratories Report*, Hull University, UK. Avaliable: www.laidback.org/~daveg/academic/labreports/rep4/SCHRBRDG.html

Glovnea, R.P. and Cretu, O.S., 2006. "Double-Cage Constant Power Continuously Variable Transmission (CP-CVT)", *Proc. of 8th Biennial Conference on Engineering Systems and Analysis ESDA*. 2006.

Goss, S., Aron, S., Deneubourg, J-L. and Pasteels, J.M., 1989 "Self-organized shortcuts in the Argentine ant," *Naturwissenschaften*, Vol.76. pp.579-581.

Gover, C.P., 1996. "What and how about quality function deployment (QFD)", *Int. J. Production Economics*. Vol.46-47. pp.575-585

Ilett, D (editor), 2004. "With global car ownership set to double, EU eyes greener transport" (Online Article) *Greenbang.com*. Available from: http://www.greenbang.com/with-global-car-ownership-set-to-double-eu-eyes-greener-transport_14284.html

Greene, D. and DeCicco, J., 2000. *Engineering-Economic Analyses of Automotive Fuel Economy Potential in the United States*, Prepared by The Oak Ridge National Laboratory for the U.S. Dept. of Energy. Report Number: ORNL/TM-2000/26

Greenwood, C., 2007. "Epicycloidal roller control (ERC) – its impact upon variator capability", *CVT-Hybrid 2007 International Congress on Continuously Variable and Hybrid Transmissions,* Yokohama, Japan. September, 2007.

Greenwood, J.A. and Tabor, D., 1957. "The Friction of Hard Sliders on Lubricated Rubber: The Importance of Deformation Losses", *Proc. Phys. Soc. LXXI 6.* pp989-1001.

Harned, J., Johnston, L. and Scharpf, G., 1969. "Measurement of Tire Brake Force Characteristics as Related to Wheel Slip (Antilock) Control System Design," *SAE Transactions*, Vol. 78, Paper 690214, pp909–925.

Harris, W., 2008. *How CVTs work*. Part of the "How Stuff Works – Auto" Articles [Online Article]. Available: http://auto.howstuffworks.com/cvt.htm.

Harris, T.A. and Kotzalas, M.N., 2007. *Essential Concepts of Bearing Technology*. CRC Press (Taylor and Francis Group): Florida, USA.

Heilich, F.W. and Shube, E.E., 1983. *Traction Drives: Selection and Application.* Marcel Dekker inc: New York, USA.

Hellman, K.H. and Heavenrich, R.M. 2001. *Light-Duty Automotive Technology and Fuel Economy Trends 1975 through 2001*, Technical Paper. Reference No. EPA420-R-01-008, Ann Arbor, MI: U.S. Environmental Protection Agency.

Hirst, W. and Moore, A.J., 1978. "Elastohydrodynamic lubrication at high pressures", *Proc. Roy. Soc. London, A,* Vol. 360, 1702, pp403-425

Hogan, T.J. and Donahue, D.R., 2008. "Continuously Variable Transmissions in the Automotive and Other Industries" *Eighth Annual University of Pittsburgh Freshman Engineering Conference*, University of Pittsburg, USA, May 2008

Holland, J.H., 1992. *Adaptation in Natural and Artificial Systems*. 2nd Edition. MIT Press: Cambridge, MA. USA.

Houser, J. and Clausing, D.P., 1988. "The house of quality", *Harvard Business Review*, May–June1988, pp63–73.

Huber, C. and Zorge, R., 1986. "A capacitive sensor responding to mass rather than level of liquid in a tank", *Journal A*, Vol. 27, No. 2, pp.87-90.

Hudlet, S., Saint-Jean, M., Guthmann, C. and Berger, J., 1998, "Evaluation of the Capacitive Force Between an Atomic Force Microscopy Tip and a Metallic Surface". *The European Physical Journal B.* Vol 2. pp.5-10.

Hurst, K. ed., 1994. *Rotary Power Transmission Design*. Berkshire, England: McGraw-Hill Book Company Europe.

Johnson, K.L., 1985. *Contact Mechanics*, Cambridge University Press, Cambridge, UK.

Kennedy, J. and Eberhart, R., 1995. "Particle Swarm Optimization", *Proceedings of IEEE International Conference on Neural Networks IV*. pp1942-1948.

Keun, I. and Gill, N., 2004. "An Introduction of a New Traction Drive Pairing with the Inner and the Outer Surface of the Spherical Rotors for Automobile Usage", *2004 International Continuously Variable and Hybrid Transmission Congress,* September 2004, University of California (UC Davis), CA, USA.

Kluger, M.A. and Fussner, D.R., 1997. "An Overview of Current CVT Mechanisms, Forces and Efficiencies", *Transmission and Driveline Systems Symposium*. SAE Paper No. 970688, in SP-1241, pp81-88.

Kluger, M.A., and Long, D.M., 1999. "An Overview of Current Automatic, Manual and Continuously Variable Transmission Efficiencies and Their Projected Future Improvements", *SAE Transactions*, 1999, SAE Paper No. 1999-01-1259, pp1-6

Koen, B., 1985. *The definition of the engineering method*, . Washington D.C., USA: The American Society of Engineering Education.

Kogut, L. and Etsion, I., 2002. "Elastic-Plastic Contact Analysis of a Sphere and a Rigid Flat". *Journal of Applied Mechanics*, Vol. 69, Iss. 4, pp657-662.

Krim, J., 1996. "Friction at the atomic scale," *Scientific American*, Vol. 275, Iss. 4.

Kuwajima, M., Koishi, M. and Sugimura, J., 2006, "Contact Analysis of Tire Tread Rubber on Flat Surface with Microscopic Roughness," *Tire Science and Technology,* TSTCA, Vol. 34, No. 4, pp237-255,

Lang, K.R., 2000, "Continuously Variable Transmissions: An Overview of CVT Research Past, Present and Future", *Homepage of Massachusetts Institute of Technology*, [Online]. Available: http://www.lasercannon.com/Murano/Files/cvt.pdf [Accessed: 9/8/2010].

Lee, A.P., Newall, J., Goto, M., Misada, Y. and Yoshihiro, O., 2004. "Experimental validation of full toroidal fatigue life", *International Continuously Variable and Hybrid Transmission Congress*, September 2004, Davis, California, USA.

Luke, P. and Olver, A.V., 1999. "A study of churning losses in dip-lubricated spur gears", *Proc Instn Mech Engrs*, Vol 213, Part G, pp337-346

Mcmanus, J.H. And Anderson, S.R., 2005. "Creating The Virtual Vehicle Model Today" *Homepage of Ricardo Company Technical Papers*, [Online]. Available from: http://www.ricardo.com/download/pdf/wave_virtual_vehicle_model.pdf [Accessed: 1/6/2009].

Merts, J.P., 2009. "The Psychology of Notation" *QFD Online* [Online article]. Available: http://www.qfdonline.com/archives/the-psychology-of-notation/ [Accessed: 10/6/2010]

Morita, T., Imayoshi, Y., Bell, C., Glovnea, R. and Sugimura, J., 2010. "Experimental study of rubber traction with concrete model surfaces", *International Tribology Congress - ASIATRIB 2010*, 5th-9th December 2010, Perth, Australia.

Myshkin, N.K., Petrokovets, M.I., and Kovalev, A.V., 2005. "Tribology of polymers: Adhesion, friction wear, and mass-transfer," *Tribology International*, Vol. *38*, pp910-921.

Newall, J., Cowperthwaite, S., Hough, M. and Lee, A., 2004. "Efficiency modelling in the full toroidal variator: Investigation into optimization of EHL contact conditions to maximize contact efficiency", *International Continuously Variable and Hybrid Transmission Congress*, September 2004, UC Davis, California, USA.

Newall, J. and Lee, A., 2003. "Measurement and prediction of spin losses in the EHL point contact of the full toroidal variator", *Proceedings of the 30th Leeds-Lyon Symposium on Tribology*, Lyon, September 2003, Tribology Series 43, D. Dowson et al. (Eds.), Elsevier, 2003, pp.769-779.

Nikas, G.K., 2002. “Fatigue Life and Traction Modelling of Continuously Variable Transmissions”. *Journal of Tribology*, Vol. 124, Iss. 4, pp689-698.

Ohno, N., 2006. “High-Pressure Behavior of Toroidal CVT Fluid For Automobile”. *Tribology International*, Vol. 40, pp233-238.

Olsson, H., Åström, K.J. Canudas de Wit, C., Gäfvert, M. and Lischinsky, P., 1998. “Friction Models and Friction Compensation” *European Journal of Control*, No.4, pp176-195.

Pacejka, H.B. and Sharp, R.S., 1991. “Shear Force Developments by Pneumatic Tires in Steady-State Conditions: A Review of Modeling Aspects”, *Vehicle Systems Dynamics*, Vol. 20, pp121–176.

Pfiffner, R. and Guzzella, L., 2001. “Optimal Operation of CVT-based Powertrains”, *Int. J. Robust Nonlinear Control*, Vol. 11, pp1003-1021

Pohl, B., Simister, M., Smithson, R. And Miller, D., 2004. “Configuration Analysis of a Spherical Traction Drive CVT/IVT”, *2004 International Continuously Variable and Hybrid Transmission Congress*, September 2004, University of California (UC Davis), California, USA.

Pugh, S., 1981. “Concept selection: a method that works”, *Proceedings of the International Conference on Engineering Design*, Rome, Italy, WDK 5 Paper M3/16, pp497-506.

Pugh, S., 1986. “Design activity models: Worldwide emergence and convergence”, *Design Studies*, Vol.7. No. 3., pp.167-173

Rashedi, E., Nezamabadi-pour, H. and Saryazdi, S., 2009. “GSA: A Gravitational Search Algorithm”, *Information Sciences*. Vol.179. pp.2232-2248

Sanda, S. And Hayakawa, K., 2005. “Traction Drive System and its Characteristics as Power Transmission”. *R&D Review of Toyota CRDL*, Vol. 40 Iss. 3, pp30-39.

Schäfer, A., Heywood, J.B. and Weiss, M.A., 2006. “Future fuel cell and internal combustion engine automobile technologies”. *Energy*, Vol.31, Iss.12, pp.2064-2087

Shah-Hosseini, H., 2008. “The intelligent water drops algorithm: a nature-inspired swarm-based optimization algorithm”, *International Journal of Bio-Inspired Computation*, Vol.1. No.1. pp.71-79.

Short, M., Pont, M.J. and Huang, Q., 2004, “Simulation of Vehicle Longitudinal Dynamics” Technical Report: ESL04-01. Embedded Systems Laboratory, University of Leicester.

Smith, M.H., Barth, E.J., Sadegh, N. and Vachtsevanos, G.J., 2004. “The Horsepower Reserve Formulation of Driveablility for a Vehicle Fitted With a Continuously Variable Transmission”, *Vehicle System Dynamics*, Vol. 41 Iss. 3, pp157-180.

Soliman, A.M.A., 2006. “Effect of Road Roughness on the Vehicle Ride Comfort and Rolling Resistance”, *SAE Technical Paper Series*, April 2006, SAE International.

Srinivas, M. and Patnaik, L.M., 1994. “Adaptive probabilities of crossover and mutation in genetic algorithms”, *IEEE Transactions on Systems, Man and Cybernetic*, Vol.24, Iss.4. pp.656-667.

Stachowiak, G.W. and Batchelor, A.W., 2005, *Engineering Tribology 3rd Edition*, Elsevier Butterworth-Heinemann: Boston, U.S.

Stone, R., 2009. “Full-toroidal variable drive transmission systems in mechanical hybrid systems - from Formula 1 to road vehicles”, *CTI Symposium and Exhibition: Innovative Automotive Transmissions,* Berlin, Germany, December 2009.

Syswerda, G., 1989. “Uniform crossover in genetic algorithms”, *Proceedings of the Third International Conference on Genetic Algorithms*. San Mateo, CA. USA. pp.2-9.

Tabor, D., 1959. “Junction growth in metallic friction: the role of combined stresses and surface contamination,” *Proceedings of the Royal Society of London. Series A,* pp378-393.

Tanaka, H., 1989, “Power Transmission of a Cone Toller Toroidal Traction Drive”, *JSME International Journal, Series III,* Vol. 32, No. 1, pp82-90

Tanaka, H., Machida, H., Hata, H. And Nakano, M., 1995. "Half-Toroidal Traction Drive Continuously Variable Power Transmissions for Automobiles". *JSME International Journal, Series C*, Vol. 38, Iss. 4, pp772-777.

Tanaka, H., Toyoda, N., Machida, H. and Imanishi, T., 2004. "Development of a 6 Power-Roller Half-Toroidal CVT-Mechanism and Efficiency", *International Continuously Variable and Hybrid Transmission Congress*, September, 2004, UC Davis, California, USA.

Tenberge, P. And Mockel, J., 2002. "Toroidal CVT with compact roller suspension". *VDI-Berichte, NR. 1709*, October 2002, pp623-637.

Terekhov, A., 1991. "Basic problems of heat calculation of gear reducers" *Proceedings of Japanese Society of Mechanical Engineers International Conference on Motion and Power Transmissions*, November 1991, pp.490-495.

Tevaarwerk, J.L. and Johnson, K.L., 1979. "The Influence of Fluid Rheology on the Performance of Traction Drives," *Journal of Lubrication Technology*, Vol. 101, no. 3, pp1015-1022.

Vij, J.K., Scaife, W.G., Calderwood, J.H., 1978, "The pressure and temperature dependence of the static permittivity and density of heptanol isomers" *Journal of Physics D: Applied Physics*, Vol.11, No.4. pp.545-560

Westlake, F. and Cameron, A., 1967. "A Study of Ultra–Thin Lubricant Films Using an Optical Technique ", *Proceedings of the Institute of Mech. Engrs*, Vol.182, Pt.3G

Whitley, D., 2001. "An Overview of Evolutionary Algorithms", *Journal of Information and Software Technology,* Vol. 43, No. 43, pp817-831.

Wicke, V., Brace, C.J. and Vaughan, N.D., 2000. "The potential for simulation of driveability of CVT vehicles", *SAE 2000 World Congress Technical Papers*, March 2000, SAE.

Wicke, V., Brace, C.J., Vaughan, N.D. and James, I., 2002. "The characterisation of driveability of CVT powertrains during acceleration transients", *VDI-Berichte Nr. 1709*, October 2002.

Witte, D.C., 1973. "Operating Torque of Tapered Roller Bearings", *Tribology Transactions*, Vol. 16, Iss. 1, pp61-67

Wright, A., Vose, M., De Jong, K., and Schmitt, L., ed., 2005. *LNCS: Foundations of Genetic Algorithms*. Springer-Verlag Berlin, Heidelberg, pg75–94

Yamamoto, T., Matsuda, K. and Hibi, T., 2001. "Analysis of the efficiency of a half-toroidal CVT", *Society of Automotive Engineers of Japan (JSAE)* Review 22, pp565-570.

Zhang, N. and Dutta-Roy, T., 2004. "An Investigation into Dynamics and Stability of a Powertrain with Half-Toroidal Type CVT", *2004 SAE International Congress on Continuously Variable and Hybrid Transmission,* September 2004, SAE Paper No. 2004-34-2886.

Zhang, Y., Chen, X., Zhang, X., Jiang, H. and Tobler, W., 2005. "Dynamic Modelling and Simulation of a Dual-Clutch Automated Lay-Shaft Transmission", *Journal of Mechanical Design*, Vol. 127, Iss. 2. pp.302-307.

Zhang, Y., Zhang, X. and Tobler, W., 2000. "A Systematic Model for the Analysis of Contact, Side Slip and Traction of Toroidal Drives", *Journal of Mechanical Design*, Vol. 122, Iss. 4, pp523-528

IX. APPENDIX

A.1 Example Calculations for Chapter 4

Figure 4-1 indicates that the first losses that must be considered are at the bearing losses and the churning losses that occur at the input shaft.

A.1.1 Bearing Losses at Input Shaft

In order to calculate the bearing losses at the input shaft, the axial forces at the input shaft must be known. From the basic laws of mechanical equilibrium, the axial forces at the input and output shafts must be equal and act in opposite directions. However since both shafts will be rotating at different speeds, the torque losses will differ. When in state of equilibrium, the axial forces acting on both shafts will be equal to the force produced by the ball screw coupling, which is simply a function of the torque acting on the output shaft. Unfortunately however, there is an added complication. To determine the linear force produced by the ball screw, the properties of the ball screw must first be determined. The ball screw must be designed carefully to ensure that there is a sufficient normal force applied to each contact. If the normal forces are insufficient then the traction coefficient will exceed the intended limit and significant wear, or even a breakdown of the contact will occur. Similarly, unnecessarily high normal forces can significantly reduce efficiency and lead to higher Hertzian pressures, which may exceed the limit of the material causing failure. The properties of the ball screw must be determined when the torque at the output is equal to its maximum, i.e. when α is equal to its maximum value of 57° (1 radian)

Since there are two contacts between the input shaft and the ball element, and only one contact between the ball element and the output shaft, it can be reasonably assumed that the highest traction coefficient will occur at the output contact. The tractive force at this contact can be easily calculated from the torque at the output shaft. At this stage the number of ball elements is irrelevant, since both the tractive force and the normal force must be divided by the total number of contacts.

The maximum output torque can be calculated from the stated input torque (120Nm), together with the overall transmission ratio of the CVT. From Equation 2.14:

$$i = \frac{\sin(102)((0.12-(0.03)\sin 57)+0.06\sin 75)}{(0.12-(0.03)\sin 57)\cos 57\sin 75-0.06\sin(102)\sin 75+(0.12-(0.03)\sin 57)(\sin(30)+\sin 57\cos 75)}$$

$$i = 2.45$$

The torque output from the engine has been established as 120Nm, hence the maximum torque at the output shaft (ignoring losses) is:

$$T_{out} = T_{in} \times i = 120 \times 2.45 = 294 \text{ Nm}$$

From Figure 2-10, the distance from the output shaft to the point of contact (r_c) is:

$$r_C = r_0 - (R_1 - R)\sin\alpha + R\sin\gamma$$

Hence the maximum tractive force at the output contact is:

$$F_{tr-C} = T_{out}/r_C = 294/[0.12-(0.03)\sin 57+0.06\sin 75] = 1927 \text{ N}$$

And so the total normal force required to ensure a maximum traction coefficient of 0.045 is simply:

$$N_C = 1927/0.045 = 42824 \text{ N}$$

Looking at a free body force diagram of a single ball element (Figure A-1), the total force required from the ball screw coupling to produce a normal force of 42824N at the output contact can be calculated thusly:

$$F_C = N_C \cos\gamma = 42824 \times \cos 75 = 11084 \text{ N}$$

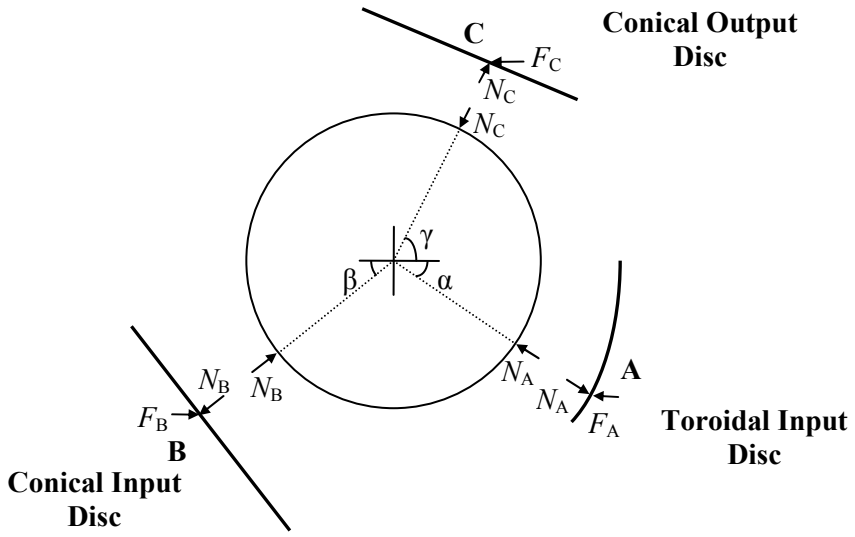


Figure A-1: Free body force diagram of single ball element

Ball screws are normally defined in terms of a ball screw lead (l), which (ignoring the inefficiencies of the ball screw itself) can be calculated from the torque applied and the linear force produced (Equation 2.17). Rearranging this:

$$l = 2\pi T / F_C = 2\pi \times 294 / 11084 = 0.166828\,\text{m}$$

Note that in reality, a ball screw lead of this length would require a very large diameter; however for the purposes of these efficiency calculations, it is assumed that this is acceptable. Now the ball screw lead has been determined, the axial force applied to the bearings at the input shaft can be calculated for the representative operating conditions (when α is set to 17°). As before, first the transmission ratio must be calculated:

$$i = \frac{\sin(62)((0.12-(0.03)\sin 17)+0.06\sin 75)}{(0.12-(0.03)\sin 17)\cos 17\sin 75-0.06\sin(62)\sin 75+(0.12-(0.03)\sin 17)(\sin(30)+\sin 17\cos 75)}$$
$$i = 1.30$$

The torque output from the engine remains the same (120Nm), hence the torque at the output shaft is now:

$$T_{out} = T_{in} \times i = 120 \times 1.30 = 156\,\text{Nm}$$

And so the axial force applied to the bearings becomes:

$$F = 2\pi T / l = 2\pi \times 156 / 0.166828 = 5859\,\text{N}$$

Now finally the torque lost to the input bearings can be calculated. From Equation 4.17:

$$T_{be-in} = 3.84 \times 10^{-5} k F_a^{1/3} \sqrt{(\omega \nu)}$$
$$T_{be-in} = 3.84 \times 10^{-5} \times 10 \times 5859^{1/3} \sqrt{((2 \times \pi \times 5000/60) \times 15)} = 0.61\,\text{Nm}$$

The power lost to bearings at the input shaft can also be calculated using the input shaft's rotational speed and the bearing torque:

$$P_{be-in} = 0.61 \times 2 \times \pi \times 5000 / 60 = 319\,\text{W}$$

A.1.2 Churning Losses at Input Shaft

Churning losses only require the dimensions of the immersed surfaces areas on the input shaft, together with their rotational speed. In order to calculate the churning losses, firstly the Reynolds number of the discs movement through the traction fluid must be determined. According to Equation 4.20, this requires the radius of the disc and the chord length of the immersed surface. The radius for both input discs can be approximated to be equal to r_0. Assuming that it is desirable to immerse the discs to the depth of the ball elements when they are closest to the input shaft (r_{min}, i.e. the value of r when α = 1, which can be calculated from

Equation 2-1), then the value of L_{im} can be determined from Figure A-2. Note that the churning losses of the ball elements about their own axis are negligible and hence ignored.

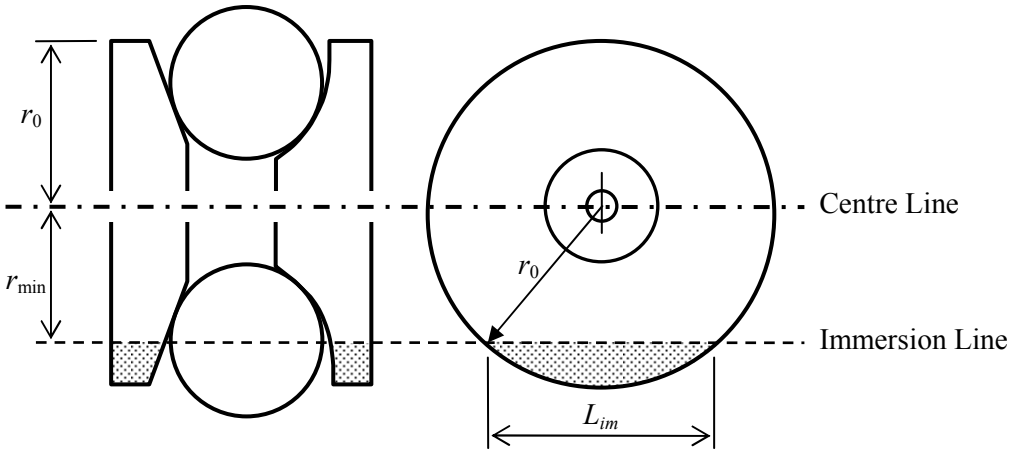


Figure A-2: Immersion depth of input discs

So:

$$R_{im} = r_{\min} = [0.12 - (0.09 - 0.06) \times \sin 1] = 94.8\,\text{mm}$$
$$L_{im} = 2\sqrt{r_0^2 - r_{\min}^2} = 2\sqrt{0.12^2 - [0.12 - (0.09 - 0.06) \times \sin 1]^2} = 147.3\,\text{mm}$$

Hence, from Equation 4.20, and assuming a kinematic viscosity of air of approximately 15Cst

$$\text{Re} = \frac{(2 \times \pi \times 5000/60) \times 94.8 \times 147.3}{15} = 487437$$

The moment coefficient can now be calculated using Equation 3.19, hence:

$$C_M = \frac{5.34 \times 10^4}{487437^{1.379}} = 0.0007654$$

The next task is to calculate the submerged surface area, which consists of the circular segment of both input discs, together with the submerged circumferential outer area. If the submerged surfaces of both discs are approximated as cylinders, and it is assumed that each disc has an additional thickness of 5mm, then the submerged surface area is simply:

$$A_{im} = 4 \times \frac{r_0^2}{2}\left(2\arccos\left(\frac{r_{\min}}{r_0}\right) - \sin\left(2\arccos\left(\frac{r_{\min}}{r_0}\right)\right)\right) + 2 \times 0.05 \times r_0 \times 2\arccos\left(\frac{r_{\min}}{r_0}\right)$$

So:

$$A_{im} = 2 \times 0.12^2\left(2\arccos\left(\frac{0.095}{0.12}\right) - \sin\left(2\arccos\left(\frac{0.095}{0.12}\right)\right)\right) + 0.2 \times 0.12 \times \arccos\left(\frac{0.095}{0.12}\right)$$
$$A_{im} = 0.0260\,\text{m}^2$$

Hence the torque lost at the input shaft due to churning becomes:

$$T_{ch-in} = \tfrac{1}{2} C_M \rho \omega_{in}{}^2 R_{im}{}^3 A_{im}$$
$$T_{ch-in} = 0.5 \times 0.0007654 \times 900 \times (2 \times \pi \times 5000/60)^2 \times 0.0948^3 \times 0.0260 = 2.09 \text{ Nm}$$

The power lost due to churning at the input shaft can be easily calculated from the rotational speed and the churning torque:

$$P_{ch-in} = 2.09 \times 2 \times \pi \times 5000/60 = 1094 \text{ W}$$

A.1.3 Input Contact Losses – Creep

Now that the losses at the input shaft have been determined, the losses at the contact point between the input discs and the ball elements can be calculated. Firstly the traction coefficient for both contact point must be calculated. It is reasonable to assume that each contact point will transfer half the total power to the ball elements. From the previous calculations, the power available at the input discs is:

$$P_{A+B} = (T_{in} - T_{be-i} - T_{ch-i})\omega_{in} = (120 - 0.61 - 2.09) \times (2 \times \pi \times 5000/60) = 61418 \text{ W}$$

The linear velocities of the contact points at the input side of each of the contacts can be easily calculated using Equations 2-2 and 2-3. These are found to be 48.91m/s and 35.98m/s for the toroidal disc and conical disc respectively. The tractive force applied to each contact at the input side is thus:

$$F_{tr-A} = \tfrac{1}{2} \times 61418/48.91 = 628 \text{ N}$$
$$F_{tr-B} = \tfrac{1}{2} \times 61418/35.98 = 854 \text{ N}$$

Using Equations 3.17-3.18, N_C can be shown to be 22639N when $\alpha = 17°$ and the toque output is 156Nm, whilst the values of N_A and N_B can be shown to be 12798N and 25577N respectively. The effective traction coefficient at each contact point can now be calculated:

$$\mu_A = F_{tr-A}/N_A = 628/12798 = 0.049$$
$$\mu_B = F_{tr-B}/N_B = 854/25577 = 0.033$$

Note that traction coefficient for the toroidal disc (μ_A) is actually higher than the design value of 0.045; however these calculations ignore any pre-compression applied to the toroidal spring in order to provide the initial traction required to turn the output disc. This will not affect the overall forces applied to the bearings, and will only serve to reduce the traction coefficient and hence the creep losses at the contacts, hence these calculations can be considered a worst-case possibility. The determination of the precise pre-compression required depends largely on the

desired drivability and control of the CVT and hence goes beyond the initial efficiency calculations. This will be discussed in more detail in later sections.

Referring to Figure 4-3, the traction coefficients found, are clearly within the linear region of the slide-roll ratio curve. This allows a very simple relationship to be determined, based on the initial slope of the traction curve (m_{tr}). Observing the curve for the traction fluid shown at 40°C:

$$m_{tr} = \frac{0.085}{0.005} = 17$$

Hence the slide-roll ratio (S) at each of the input contact points is:

$$S_A = 0.049/17 = 0.00289$$
$$S_B = 0.033/17 = 0.00194$$

From Equation 4.4

$$u_{cr} = u_1\left[1 - \frac{(2-S)}{(2+S)}\right]$$

The values of $u_{in\text{-}A}$ and $u_{in\text{-}B}$ have already been determined as 48.91m/s and 35.98m/s respectively. Hence the magnitude of creep at each contact is:

$$u_{cr-A} = 48.91 \times \left[1 - \frac{(2-0.00289)}{(2+0.00289)}\right] = 0.1377 \text{ m/s}$$
$$u_{cr-B} = 35.98 \times \left[1 - \frac{(2-0.00194)}{(2+0.00194)}\right] = 0.0697 \text{ m/s}$$

Since the ball can obviously only rotate at a single speed, it is assumed that the rotational speed of the ball will be the average of each of the speeds of the contact points, i.e. from Equations 2.4 and 2.5:

$$\omega_{ball} = \frac{1}{2}\left[\frac{u_{ball-A}}{R\sin(\alpha-\lambda)} + \frac{u_{ball-B}}{R\sin(\beta+\lambda)}\right] = \frac{1}{2}\left[\frac{u_{in-A}-u_{cr-A}}{R\sin(\alpha-\lambda)} + \frac{u_{in-B}-u_{cr-B}}{R\sin(\beta+\lambda)}\right]$$

Using Equation 2.7, λ was found to be -19.15°, hence:

$$\omega_{ball} = \frac{1}{2}\left[\frac{48.91-0.1377}{0.06\sin(17+19.15)} + \frac{35.98-0.0697}{0.06\sin(45-19.15)}\right] = 1372 \text{ rad/s}$$

If creep was ignored, the ball element should be rotating at 1375.4rad/s, hence in simple terms, creep at the input contacts has caused a reduction in the ball element's rotational speed of 3.3rad/s, i.e. a reduction in speed of less than 1%.

The exact power loss can be easily calculated by multiplying the tractive force acting on each contact by the creep velocities. Hence:

$$P_{cr-in} = (0.1377 \times 628) + (0.0697 \times 854) = 146.0\,\text{W}$$

A.1.4 Input Contact Losses – Spin

Spin losses are perhaps the most complex and time consuming to calculate. Using Hertzian theory discussed in the previous section, together with Equations 4.10-4.11, the following properties were derived for contacts *A* (toroidal disc-ball) and *B* (conical input disc-ball):

Table A-1: Calculated properties of input contacts

Parameter	*Contact A*	*Contact B*
Central film thickness h_c	1.34×10^{-6}m	1.10×10^{-6}m
Effective contact semi-width c	0.0020m	0.0019m
Contact semi-width in rolling direction b	0.0013m	0.0017m
Hertzian Pressure P_m	0.255GPa	0.512GPa
Average fluid shear modulus G	12.4MPa	15.7MPa

The spin velocities of both contacts are relatively simple to calculate from the geometry of the input discs, together with Equations 4.5-4.7:

$$\omega_{sp-A} = \left[\omega_{in} \sin(90-\alpha)\right] - \left[\omega_{ball} \sin(\alpha - 90 + \lambda)\right]$$
$$\omega_{sp-B} = \left[\omega_{in} \sin(90-\beta)\right] - \left[\omega_{ball} \sin(\beta - 90 - \lambda)\right]$$

Hence:

$$\omega_{sp-A} = \left[(2 \times \pi \times 5000/60) \times \sin(90-17)\right] - \left[1372.1 \times \sin(17-90-19.15)\right]$$
$$= 1841.5 rad/s$$
$$\omega_{sp-B} = \left[(2 \times \pi \times 5000/60) \times \sin(90-45)\right] - \left[1372.1 \times \sin(45-90+19.15)\right]$$
$$= 968.4 rad/s$$

The dimensionless spin factor requires the addition of the tangential speeds for both sides of the contact as shown in Equation 4.9. This is the equivalent of taking twice the velocity at the input side and subtracting the creep velocity. i.e.

$$\psi_A = \frac{2 \times 1841.5 \times 0.002}{(2 \times 48.91) - 0.1377} = 0.0754$$
$$\psi_B = \frac{2 \times 968.4 \times 0.0019}{(2 \times 35.98) - 0.0697} = 0.0537$$

Assuming a Poisson's ratio of 0.3 (steel), the spin torque per ball element is thus:

$$T_{sp-A} = \frac{32(2-0.3)}{9(3-0.6)} \times 12.4 \times 10^6 \times 0.0013^3 \times 0.0754 = 0.0052 \text{ Nm}$$
$$T_{sp-B} = \frac{32(2-0.3)}{9(3-0.6)} \times 15.7 \times 10^6 \times 0.0017^3 \times 0.0537 = 0.0104 \text{ Nm}$$

As expected, the spin torque is particularly small. To calculate the power lost per contact to spin the spin torques are simply multiplied by the spin velocities:

$$P_{sp-A} = 0.0052 \times 1841.5 = 9.58 \text{ W}$$
$$P_{sp-B} = 0.0104 \times 968.4 = 10.07 \text{ W}$$

These values are still relatively low, however unlike creep, which is largely independent on the number of contact points, spin losses occur at every contact hence must be multiplied by the number of ball elements. Individually, it is more useful to determine the tangential force lost across the contact acting against the direction of traction. Assuming four ball elements, this can be calculated using the tangential speeds:

$$F_{sp-A} = 4 \times 9.58 / 48.91 = 0.78 \text{N}$$
$$F_{sp-B} = 4 \times 10.07 / 35.98 = 1.12 \text{N}$$

This give a total power lost to spin for the input contacts of 78.6W.

A.1.5 Cage Losses

The cage losses are relatively simple to calculate, requiring no additional information to be calculated. There is however an additional complication in determining a suitable value of traction coefficient for the contact between the cage and the ball elements. Assuming a simple cage is used, the nature of the contact will be pure sliding, and hence from Figure 4-3 it might be assumed that the traction coefficient would be in the region of 0.1-0.12 for an operating temperature of approximately 40°C. This assumes a very poorly designed cage and would lead to exceptionally high losses. A more sensible approach would be to have a deep-groove cage race-way, which would significantly reduce the Hertzian pressure of the contact, possibly making the nature of the contact hydrodynamic, rather than elastohydrodynamic. This would greatly reduce the effective traction coefficient. Further reductions could be made through he

use of low friction coatings on the contacting surface of the cage. All of these factors make it exceptionally difficult to determine a precise value of traction coefficient. An optimistic value would be approximately 0.02. The torque of the ball element lost to the cage is hence:

$$T_{ca} = \frac{R \times (T_{in} - T_{be-in} - T_{ch-in}) \times \mu}{r_0 - (R_1 - R)\sin\alpha} = \frac{0.06 \times (120 - 0.61 - 2.09) \times 0.020}{0.12 - (0.09 - 0.06) \times \sin 17} = 1.27 \text{ Nm}$$

Whilst the power lost to the cage is:

$$P_{ca} = T_{ca} \times \omega_{ball} = 1.27 \times 1372.1 = 1742 \text{ W}$$

A.1.6 Output Contact Losses – Creep

Before calculating the contact losses between the ball elements and the output disc, it is necessary to establish the tangential force and speed at the ball side of the contact. The rotational speed of the ball including creep losses at the input contacts has already been determined; hence the tangential speed output contact can be calculated using Equation 2.8:

$$u_{ball-C} = \omega_{ball}[R\sin(\gamma + \lambda)] = 1372.1 \times [0.06 \times \sin(75 - 19.15)] = 68.13 \text{ m/s}$$

The tangential force applied to the ball side of the contact is a little more difficult to calculate:

$$F_{tr-C} = \frac{T_{ball} - T_{ca}}{R\sin(\gamma + \lambda)} = \frac{[(F_{tr-A} - F_{sp-A})R\sin(\alpha - \lambda)] + [(F_{tr-B} - F_{sp-A})R\sin(\beta + \lambda)] - T_{ca}}{R\sin(\gamma + \lambda)}$$

Hence:

$$F_{tr-C} = \frac{[(628 - 0.78) \times 0.06 \times \sin(17 + 19.15)] + [(854 - 1.12) \times 0.06 \times \sin(45 - 19.15)] - 1.27}{0.06 \times \sin(75 - 19.15)}$$

$$= 870.9 \text{N}$$

The normal force at contact C has already been determined; hence the traction coefficient is simply:

$$\mu_C = \frac{F_{tr-C}}{N_C} = \frac{870.9}{22639} = 0.0385$$

The slide-roll ratio is thus 0.0385/17 = 0.00226, hence the magnitude of creep is:

$$u_{cr-C} = 68.13 \times \left[1 - \frac{(2 - 0.00226)}{(2 + 0.00226)}\right] = 0.153 \text{ m/s}$$

From this, and Equation 2.9, the rotational speed of the output shaft (ω_{out}) can be calculated:

$$\omega_{out} = \frac{u_{ball-C} - u_{cr-C}}{r_0 - (R_1 - R)\sin\alpha + R\sin\gamma} = \frac{68.13 - 0.153}{0.12 - 0.03 \times \sin 17 + 0.06 \times \sin 75} = 401.8\,\text{rad/s}$$

The ideal transmission ratio (i), which can be calculated from Equation 2.14, was found to be 1.296 at $\alpha = 17°$. Hence if the creep at all the contacts was ignored, the rotational speed of the output shaft should be:

$$\omega_{\substack{out \\ ideal}} = \frac{\omega_{in}}{i} = \frac{(2 \times \pi \times 5000/60)}{1.296} = 404.0\,\text{rad/s}$$

Hence, slip at the all the contacts has only caused a reduction in the output shaft's rotational speed of 2.2rad/s, or 0.5%. This is perhaps one of the reasons it is typically ignored in efficiency calculations. The actual power lost can be found simply from the creep velocity and the tractive force, i.e. 0.153 x 870.9 = 133.2W.

A.1.7 Output Contact Losses – Spin

As before, using Hertzian theory and Equations 4.10-4.11, the following properties were derived for the output contact *C*:

Table A-2: Calculated properties of output contacts

Parameter	*Contact C*
Central film thickness h_c	1.69×10^{-6}m
Effective contact semi-width c	0.0022m
Contact semi-width in rolling direction b	0.0019m
Hertzian Pressure P_m	0.364GPa
Average fluid shear modulus G	15.0MPa

From the geometry of the output disc, the spin velocity is:

$$\omega_{sp-C} = \left[\omega_{ball}\sin(90 - \gamma - \lambda)\right] - \left[\omega_{out}\sin(-\lambda)\right]$$
$$\omega_{sp-C} = \left[1372.1 \times \sin(90 - 75 - 19.15)\right] - \left[401.8 \times \sin(-75)\right]$$
$$= 1158.5 rad/s$$

As before:

$$\psi_C = \frac{2 \times 1158.5 \times 0.0022}{(2 \times 68.13) - 0.153} = 0.0375$$

$$T_{sp-C} = \frac{32(2-0.3)}{9(3-0.6)} \times 15.0 \times 10^6 \times 0.0019^3 \times 0.0375 = 0.00972 \text{ Nm}$$

$$P_{sp-C} = 0.00972 \times 1158.5 = 11.26 \text{ W}$$

Again, assuming four ball elements, the total power lost to spin in the output contact is 45.0W.

A.1.8 Churning Losses at Input Shaft

The churning losses for the output shaft may seem a little more difficult to calculate. The immersion radius remains the same however, and the only main difference is the immersed area and chord length.

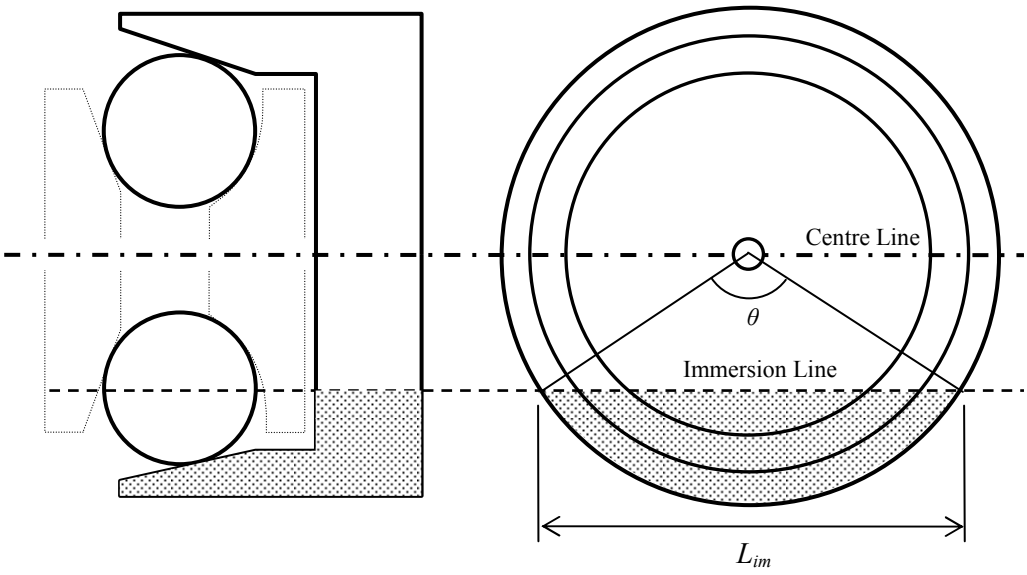


Figure A-3: Immersion depth of output discs

As before, R_{im} = 94.8mm. Ignoring the small addition created by the thickness of the disc, *Lim* can be approximated as:

$$L_{im} = 2\sqrt{(r_0 + R\sin\gamma)^2 - r_{\min}^2} = 2\sqrt{[0.12 + (0.06 \times \sin 75)]^2 - [0.12 - (0.03 \times \sin 1)]^2}$$

$$L_{im} = 0.3013\text{m} = 301.3\text{mm}$$

As shown previously:

$$\text{Re} = \frac{401.8 \times 94.8 \times 301.3}{15} = 765114$$

$$C_M = \frac{5.34 \times 10^4}{765114^{1.379}} = 0.000411$$

The submerged surface area c an be approximated to a cylinder with a single end. Hence the total area consists of twice the circular segment of the end of the cylinder and the circumferential outer area. Hence:

$$A_{im} = 2 \times \left[\left(\frac{\theta(r_0 - r_{\min})}{\tan(90 - \gamma)} \times (r_0 + R\sin\gamma) \right) + \left(\frac{r_0 + R\sin\gamma}{2} (\theta - \sin\theta) \right) \right]$$

Where θ is the arc angle in radians shown in Figure A-3, which can be calculated as:

$$\theta = 2\arccos\left(\frac{r_{\min}}{r_0 + R\sin\gamma} \right) = 2.02 \text{ rads}$$

Using these equations, A_{im} was found to be 0.103m^2, which, as expected, is significantly higher than the input discs immersed area. The torque lost at the output shaft due to churning becomes:

$$T_{ch-out} = 0.5 \times 0.000411 \times 900 \times 401.8^2 \times 0.0948^3 \times 0.103 = 2.62 \text{ Nm}$$

The power lost due to churning at the output shaft can now be calculated:

$$P_{ch-out} = 2.62 \times 401.8 = 1053 \text{ W}$$

A.1.9 Bearing Losses at Output Shaft

The final losses according to Figure 4-1 are the bearing losses at the output shaft. Since axial force and the rotational speed of the output shaft are already known, these losses can be calculated directly:

$$T_{be-out} = 3.84 \times 10^{-5} \times 10 \times 5859^{1/3} \sqrt{(401.8 \times 15)} = 0.537 \text{ Nm}$$

The power lost to bearings at the input shaft can also be calculated using the input shaft's rotational speed and the bearing torque:

$$P_{be-out} = 0.537 \times 401.8 = 216 \text{W}$$

A.2 Creation of Contoured Surfaces for the Assessment of Algorithms

In order to test the effectiveness of a number of different search algorithms, a 2-dimensional array of values of dimensions 30x30 was randomly created, with values ranging from 0 to 1. This is designed to represent a typical complicated function with two inputs, with the aim of finding the largest value, and the location at which it occurs. The surfaces were created using random number functions built into Microsoft Excel.

The simple, single peak terrain was created by layering two different surfaces together, one consisting of low-level random noise, and the other to generate a clearly defined dominant peak.

A.2.1 Noise generation

Low level noise is created by initially creating a random pattern in one-dimension. Starting from reference point of 0, each successive step has a probability of either going up or down by a random amount (*RAND*), or staying at the same level, i.e.

$$x_i = x_{i-1} \pm RAND(0 \rightarrow 1)$$

The result of this is a single dimensional, as shown below:

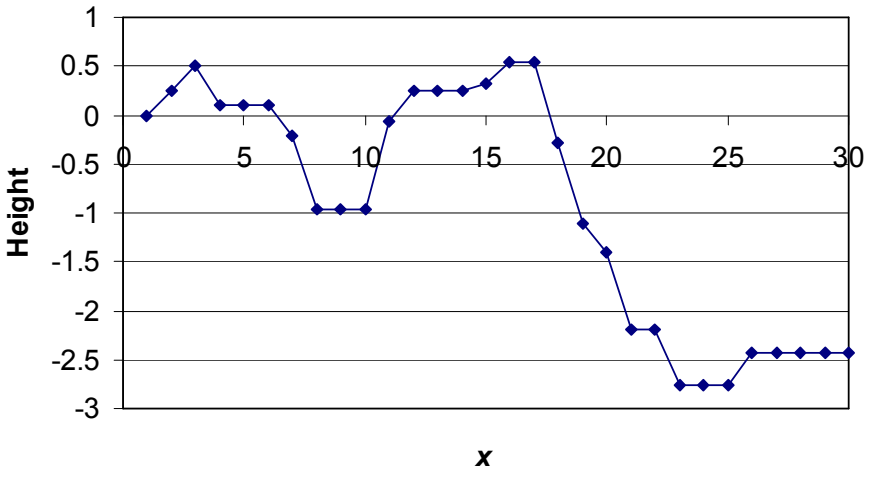


Figure A-4: Single dimension of randomly created, step-noise

Additional lines are then created running parallel to this to create a two-dimensional array. Each point incorporates the heights of all preceding adjacent points to ensure that there is no discontinuity within the surface, i.e.

$$(x_i, y_j) = AVG[(x_{i-1}, y_{j-1}); (x_i, y_{j-1}); (x_{i-1}, y_j)] \pm RAND(0 \rightarrow 1)$$

This results in the following surface:

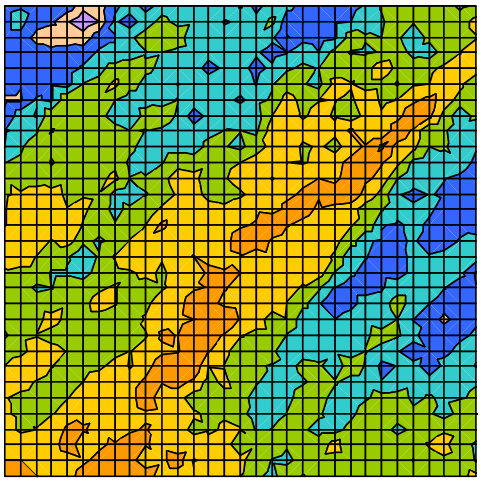


Figure A-5: Randomly generated 2-dimensinoal noise

A.2.2. Single peak generation

In order to test the effectiveness of algorithms on simple terrains, a single peak array is created. This is created by assigning a random point, approximately central to the array, to be the highest point. Surrounding points then 'fall-off' from this point at a random rate until the edge of the array. This results in the terrain shown:

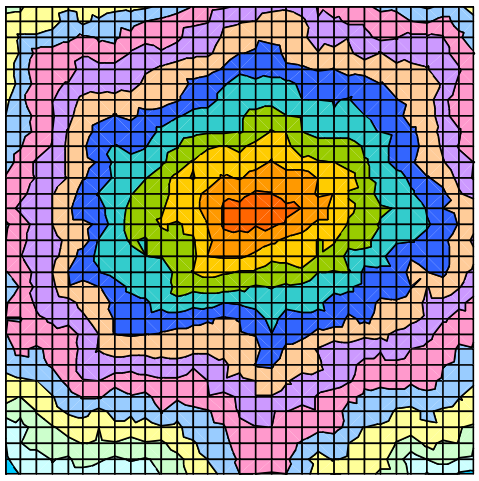


Figure A-6: Randomly created single peak, simple terrain

This surface is considered too simple, and hence is layered with the noise created previously. The amplitude of the noise is significantly lower than the amplitude of the peak, ensuring that

only one peak remains and the overall shape is unaffected. The result of layering these two surfaces is shown below.

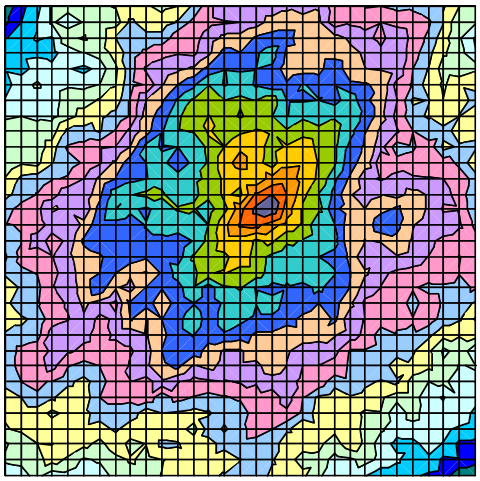


Figure A-7: Randomly created single peak, simple terrain with noise

A.2.3 Complex terrain

Complex, multi-peak terrain is created by layering multiple single peak terrains together, the result of which is shown here:

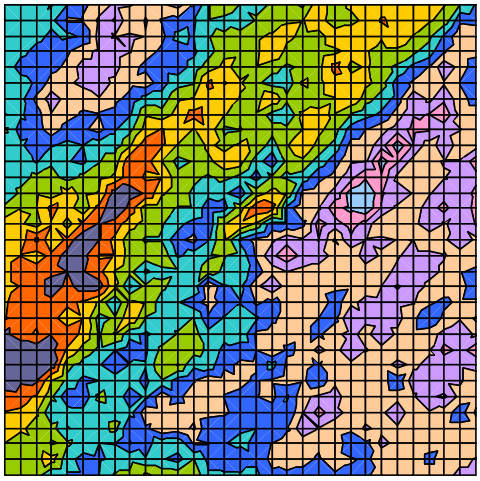


Figure A-8: Complex, multiple peak terrain

All terrains are normalised by subtracting the minimum value found from all array values then dividing by the maximum height difference. This ensures that largest value found is 1, whilst the lowest point is 0.

A.3 Derivation of Volumetric Calculations of Key Components

A.3.1 Conical Input Disc

The mass the conical input disc can be calculated from the figure shown below:

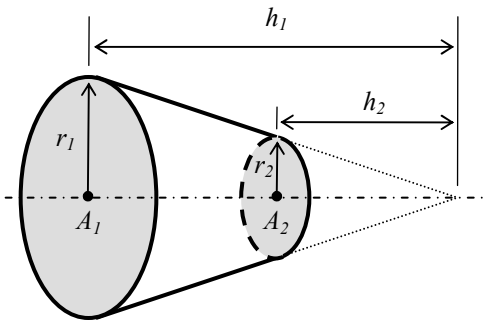


Figure A-9: Volumetric calculation of conical input disc

Hence the total volume is:

$$V_{ci} = \frac{1}{3} A_1 h_1 - \frac{1}{3} A_2 h_2$$

Or, in terms of the radii:

$$V_{ci} = \frac{\pi}{3}\left[r_1^2 h_1 - r_2^2 h_2\right]$$

From previous equations the values of r_1, r_2, h_1 and h_2 are known:

$$r_1 = r_0 - R\sin\beta \qquad r_2 = r_0 - (R_1 - R)\sin\alpha - R\sin\beta$$
$$h_1 = r_1 \tan\beta \qquad h_2 = r_2 \tan\beta$$

Hence:

$$V_{ci} = \tfrac{1}{3}\pi \tan\beta\left([r_0 - R\sin\beta]^3 - [r_0 - (R_1 - R)\sin\alpha - R\sin\beta]^3\right)$$

A.3.2 Conical Output Disc

The conical output disc can be calculated in a similar method by subtracting the volume calculated from a simple cylinder. Assuming a 10mm thickness is added to cylinder:

$$V_{co} = \pi\tan\gamma\left((r_0 + R\sin\gamma + 0.01)^2 (R_1 - R)\sin\alpha - \tfrac{1}{3}(r_0 + R\sin\gamma)^3 + \tfrac{1}{3}[r_0 - (R_1 - R)\sin\alpha + R\sin\gamma]^3\right)$$

A.3.3 Toroidal Input Disc

The volume of the toroidal input disc is significantly more complicated to calculate. In order to calculate this, Figure A-10 shows a section of the toroidal disc plotted on an *x*-*y* axis.

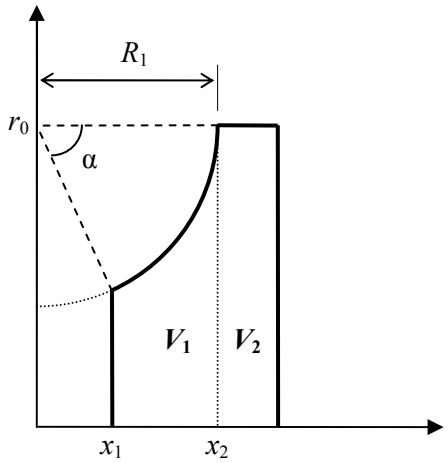


Figure A-10: Toroidal surface plotted on *x*-*y* axis

Looking only at V_1, the equation of the curve (from the standard equation for a circle) is:

$$y = r_0 - \sqrt{R_1^2 - x^2}$$

From Figure A-10, the values of x_1 and x_2 are simply $R_1\cos\alpha$ and R_1 respectively. The volume of a curve rotated about *x*-axis can be found by taking the integral of the function of y^2:

$$V_1 = \pi \int_{x_1}^{x_2} [y^2] \cdot dx$$

From the equation for *y*:

$$y^2 = \left(r_0 - \sqrt{R_1^2 - x^2}\right)^2 = r_0^2 + R_1^2 - x^2 - 2r_0\sqrt{R_1^2 - x^2}$$

Now, let:

$$x = R_1 \sin u$$

Then:

$$x^2 = R_1^2 \sin^2 u$$

And:

$$\tfrac{dx}{du} = R_1 \cos u$$

Hence:

$$y^2 = r_0^2 + R_1^2 - R_1^2 \sin^2 u - 2r_0\sqrt{R_1^2 - R_1^2 \sin^2 u}$$

Simplifying:

$$y^2 = r_0^2 + R_1^2 - R_1^2 \sin^2 u - 2r_0 R_1 \cos\ u$$

New limits when $x_2 = R_1; u_2 = \sin^{-1}\frac{R_1}{R_1} = \sin^{-1} 1 \Rightarrow u_2 = \frac{\pi}{2}$

Similarly, when $x_1 = R_1 \cos\alpha; \sin u_1 = \cos\alpha \Rightarrow u_1 = \left(\frac{\pi}{2} - \alpha\right)$

Now, given that $dx = \frac{dx}{du} du$, it can be stated that:

$$dx = R_1 \cos u \cdot du$$

Combining previous equations:

$$V_1 = \pi \int_{\frac{\pi}{2}-\alpha}^{\frac{\pi}{2}} [r_0^2 + R_1^2 - R_1^2 \sin^2 u - 2r_0 R_1 \cos\ u] R_1 \cos u \cdot du$$

Simplifying:

$$V_1 = \pi \int_{\frac{\pi}{2}-\alpha}^{\frac{\pi}{2}} [R_1^3 \cos^3 u - 2r_0 R_1^2 \cos^2 u + r_0^2 R_1 \cos u] \cdot du$$

Expanding using trigonometry power reducing formulae:

$$V_1 = \pi \int_{\frac{\pi}{2}-\alpha}^{\frac{\pi}{2}} [\tfrac{3}{4} R_1^3 \cos u + \tfrac{1}{4} R_1^3 \cos 3u - r_0 R_1^2 - r_0 R_1^2 \cos 2u + r_0^2 R_1 \cos u] \cdot du$$

Hence:

$$V_1 = \pi [\tfrac{3}{4} R_1^3 \sin u + \tfrac{1}{12} R_1^3 \sin 3u - u r_0 R_1^2 - \tfrac{1}{2} r_0 R_1^2 \sin 2u + r_0^2 R_1 \sin u]_{\frac{\pi}{2}-\alpha}^{\frac{\pi}{2}}$$

So:

$$V_1 = \pi [\tfrac{3}{4} R_1^3 - \tfrac{1}{12} R_1^3 - \tfrac{\pi}{2} r_0 R_1^2 + r_0^2 R_1]$$
$$- \pi \left[\tfrac{3}{4} R_1^3 \cos\alpha - \tfrac{1}{12} R_1^3 \cos 3\alpha - (\tfrac{\pi}{2} - \alpha) r_0 R_1^2 - \tfrac{1}{2} r_0 R_1^2 \sin 2\alpha + r_0^2 R_1 \cos\alpha\right]$$

Finally:

$$V_1 = \pi [\tfrac{2}{3} R_1^3 + r_0^2 R_1 - \tfrac{3}{4} R_1^3 \cos\alpha + \tfrac{1}{12} R_1^3 \cos 3\alpha - \alpha r_0 R_1^2 + \tfrac{1}{2} r_0 R_1^2 \sin 2\alpha - r_0^2 R_1 \cos\alpha]$$

An additional thickness is also required to strengthen the edge of the disc (V_2 in Figure A-10). Assuming a thickness of 10mm:

$$V_2 = 0.01 \times \pi r_0^2$$

Hence the total volume of the toroidal disc is:

$$V_{tr} = V_1 + V_2$$
$$V_{tr} = \pi [\tfrac{1}{12} R_1^3 (8 - 9\cos\alpha + \cos 3\alpha) + r_0 R_1^2 (\tfrac{1}{2} \sin 2\alpha - \alpha) + r_0^2 (R_1 - R_1 \cos\alpha + 0.01)]$$

Where α is measured in radians.

In comparison, the volume of the ball elements is relatively simple to calculate:

$$V_{be} = n\tfrac{4}{3}\pi R^3$$

Where n is the number of ball elements (typically assumed to be four).

The total mass can now be calculated using the density of the disc and ball material (ρ):

$$m = \rho\left(V_{ci} + V_{co} + V_{tr} + V_{be}\right)$$

A.4 Example Calculations for Chapter 5

In order to better demonstrate how each of the stated parameters are calculated, the original dimensions shown in Table 9 are used ($\beta = 45°$; $\gamma = 75°$; $R = 0.06$m; $R_1 = 0.09$m; $r_0 = 0.12$m).

A.4.1 Normalised Transmission Ratio

From Equation 2.14, the transmission ratio when $\alpha = 0$ is simply:

$$i_{\min} = \frac{\sin 45 \times (0.12 + 0.06 \times \sin 75)}{0.12 \times \sin 75 - 0.06 \times \sin 45 \sin 75 + 0.12 \times \sin 30}$$
$$i_{\min} = 0.933$$

Assuming a maximum value of α of 57° as before, the maximum transmission ratio is:

$$i_{\max} = \frac{\sin 102 \times ((0.12 - 0.03 \times \sin 57) + 0.06 \times \sin 75)}{(0.12 - 0.03 \times \sin 57)\cos 57 \sin 75 - 0.06 \times \sin 102 \sin 75 + (0.12 - 0.03 \times \sin 57)(\sin 30 + \sin 57 \cos 75)}$$
$$i_{\max} = 2.452$$

Hence the normalised ratio:

$$i_{norm} = \frac{i_{\max}}{i_{\min}} = \frac{2.452}{0.933} = 2.630$$

A.4.2 Torque Capacity

Hertzian calculations on the geometry of the conical input disc indicated that the reduced contact radius (R_σ) is 0.0446m. Assuming a maximum Hertzian pressure of 2.5GPa, and a contact modulus ($\overline{E}$) of 114GPa, the maximum normal force per contact at the conical input disc is:

$$N_B = \frac{P_0^{\,3} \pi^3 R_\sigma^{\,2}}{6\overline{E}^2} = \frac{(2.5 \times 10^9)^3 \times \pi_3 \times 0.0446^2}{6 \times 114 \times 10^9} = 12290\,\text{N}$$

The highest Hertzian pressure will occur when the largest torque is applied to the output shaft, hence the torque capacity must be calculated when α is at its maximum value of 57°. Assuming four contact points, the total normal force available for traction at the output contact is:

$$N_C = 4 \times 12290 \times \frac{(\cos 45 \tan 57 + \sin 45)}{(\sin 75 + \cos 75 \tan 57)} = 64940\,\text{N}$$

And so the maximum torque output becomes:

$$T_{out} = 64940 \times 0.045 \times [0.12 - (0.03 \times \sin 57) + 0.06 \times \sin 75 = 446\,\text{Nm}$$

Whilst the maximum allowable input torque is:

$$T_{in} = \frac{T_{out}}{i} = \frac{446}{2.452} = 181\,\text{Nm}$$

A.4.3 Mass Calculations

Assuming a maximum value of α of 57° (1 radian), from Equation 5.7, the volume of the conical input disc is:

$$V_{ci} = \tfrac{1}{3}\pi \tan 45[(0.12 - 0.06 \times \sin 45)^3 - (0.12 - 0.03 \times \sin 57 - 0.06 \times \sin 45)^3]$$
$$V_{ci} = 0.0003381\,\text{m}^3$$

Similarly, from Equation 5.8, the volume of the conical output disc is:

$$V_{co} = \pi \tan 75[(0.12 + 0.06 \times \sin 75 + 0.01)^2 \times 0.03 \times \sin 57$$
$$-\tfrac{1}{3}(0.12 + 0.06 \times \sin 75)^3 + \tfrac{1}{3}(0.12 - 0.03 \times \sin 57 + 0.06 \times \sin 75)^3]$$
$$V_{co} = 0.0023380\,\text{m}^2$$

For the toroidal volume calculation, α must be taken in radians:

$$V_{tr} = \pi[\tfrac{0.09^3}{12} \times (8 - 9 \times \cos 1 + \cos 3) + (0.12 \times 0.09^2 \times (\tfrac{1}{2}\sin 2 - 1))$$
$$+ 0.12^2 \times (0.09 - 0.09 \times \cos 1 + 0.01)]$$
$$V_{tr} = 0.00106856\,\text{m}^2$$

And finally, assuming four ball elements:

$$V_{be} = 4 \times \tfrac{4}{3} \times \pi \times 0.06^3 = 0.0036191\ \text{m}^2$$

Hence the total mass of the key components (assuming a material density of 7700kg/m^3) is:

$$m = 7700 \times (0.0003381 + 0.0023380 + 0.00106856 + 0.0036191) = 56.7\,\text{kg}$$

A.4.4 Significant length and diameter

From Equation 5.9 there are two possible values of x_1 depending on the geometry and dimensions of the CVT, hence both must be calculated:

$$x_1 = \min\begin{pmatrix} -R \\ ((R_1 - R)\sin 57)(\tan\beta - \tan\gamma) + R\cos\gamma \end{pmatrix}$$
$$x_1 = \min\begin{pmatrix} -0.06 \\ (0.03 \times \sin 57) \times (\tan 45 - \tan 75) + (0.06 \times \cos 75) \end{pmatrix} = \min\begin{pmatrix} -0.0600 \\ -0.0532 \end{pmatrix}$$

Hence x_1 = -0.0600m. Whilst x_2 is simply:

$$x_2 = R\cos\gamma + (R_1 - R)\sin 57 \tan\gamma + R(\cos 57 - \cos\gamma) + R_1(1 - \cos\alpha)$$
$$x_2 = (0.06 \times \cos 75) + (0.03 \times \sin 57 \times \tan 75) + (0.06 \times (\cos 57 - \cos 75)) + 0.09 \times (1 - \cos 57)$$

$$x_2 = 0.1676\,\text{m}$$

And thus the significant length of the key components is:

$$l = 0.1676 - (-0.06) = 0.2276\,\text{m}$$

Whilst the significant diameter of the key components is:

$$d = r_0 + R$$
$$d = 0.12 + 0.06 = 0.18\,\text{m}$$

A.4.5 Power Variation

In order to prove the assumptions stated, it is useful to show both methods of calculations. The calculation of *dx* is relatively simple. Taking an arbitrary α value of 22.9° (0.4 radians):

$$dx = 0.03 \times (1 + \sin 22.9 \times \tan 45 - \cos 22.9) = 0.01404\,\text{m}$$

From previous calculations the maximum value of T_{in} is 181Nm, whilst the transmission ratio at α = 22.9° is 1.430, which corresponds to a torque output of 260Nm. From previous calculations, the ball screw lead required for these dimensions to ensure a maximum traction coefficient of 0.045 is 0.167m. Hence:

$$F_C = \frac{2\pi T}{l} = \frac{2 \times \pi \times 260}{0.167} = 9791\,\text{N}$$

This corresponds to a normal force at the output contact (N_C) of 37830N. Using the equations stated previously, it can be shown that to balance this force, N_A must be 20412N, hence:

$$F_A = N_A \cos\alpha = 20412 \times \cos 22.9 = 18801\,\text{N}$$

The same process can be repeated for each value of α, as shown in Table A-3.

Table A-3: Calculations for linear regression in terms of dF_A

α [rads]	α [°]	F_A N	dF_A N	dx [m]	dx^2	dF_A^2	$dx dF_A$
0	0.0	17444	0	0	0	0	0
0.2	11.5	18195	751	0.0066	0.000043	563894	4.925
0.4	22.9	18801	1357	0.0141	0.000197	1841453	19.067
0.6	34.34	19149	1705	0.0222	0.000492	2906927	37.815
0.8	45.8	19005	1562	0.0306	0.000938	2438937	47.819
1	57.3	17937	493	0.0390	0.001524	243330	19.255
		Σ	5868	0.1124	0.003194	7994541	128.881

The correlation coefficient is thus:

$$r = \frac{n\sum\left(dx \times dF_A\right) - \sum dx \sum dF_A}{\sqrt{\left(n\sum\left(dx^2\right) - \left(\sum dx\right)^2\right) \times \left(n\sum\left(dF_A{}^2\right) - \left(\sum dF_A\right)^2\right)}}$$

$$r = \frac{6 \times 128.881 - 0.1124 \times 2868}{\sqrt{(6 \times 0.003194 - 0.1124^2) \times (6 \times 7994541 - 5858^2)}} = 0.382$$

This can be simplified significantly through the use of the transmission ratio, which is easier to calculate. Once again, using $\alpha = 22.9°$, the transmission ratio was found to be 1.430. This must now be multiplied by the function stated in Equation 5.13.

$$F_A \propto i \cos\alpha \frac{(\sin\gamma - \cos\gamma \tan\beta)}{(\cos\alpha \tan\beta + \sin\alpha)}$$

$$kF_A = 1.430 \times \cos 22.9 \times \frac{(\sin 75 - \cos 75 \times \tan 45)}{(\cos 22.9 \times \tan 45 + \sin 22.9)}$$

$$kF_A = 0.711$$

Where k is an arbitrary constant, which has no significance in this context.

As before, this can be repeated for the remainder of the values of α, as shown in Table A-4.

Table A-4: Calculations for linear regression in terms of dkF_A

α [rads]	α [°]	kF_A N	dkF_A N	dx [m]	dx^2	$dkF_A{}^2$	$dxdkF_A$
0	0.0	0.659	0	0	0	0	0
0.2	11.5	0.688	0.028	0.0066	0.000043	0.00081	0.00186
0.4	22.9	0.711	0.051	0.0141	0.000197	0.00263	0.000721
0.6	34.34	0.724	0.064	0.0222	0.000492	0.00415	0.001430
0.8	45.8	0.718	0.059	0.0306	0.000938	0.00349	0.001808
1	57.3	0.678	0.019	0.0390	0.001524	0.00035	0.000728
		Σ	0.2218	0.1124	0.003194	0.01143	0.004872

Hence:

$$r = \frac{n\sum\left(dx \times dF_A\right) - \sum dx \sum dF_A}{\sqrt{\left(n\sum\left(dx^2\right) - \left(\sum dx\right)^2\right) \times \left(n\sum\left(dF_A{}^2\right) - \left(\sum dF_A\right)^2\right)}}$$

$$r = \frac{6 \times 0.004872 - 0.1124 \times 0.2218}{\sqrt{(6 \times 0.003194 - 0.1124^2) \times (6 \times 0.01143 - 0.2218^2)}} = 0.382$$

A.5 Coding of Physical Dimensional Constraints

```
Do Until Collision = False
  beta = (Rand.Next() * 80) + 5
  gamma = (Rand.Next() * 80) + 5
  R = (Rand.Next() * 0.1) + 0.05
  R1 = (Rand.Next() * 0.1) + 0.05
  r0 = (Rand.Next() * 0.1) + 0.05

  Collision = CollisionDetection(beta, gamma, R, R1, r0)
Loop


Public Function CollisionDetection(ByRef beta As Single, ByRef gamma
As Single, ByRef R As Single, ByRef R1 As Single, ByRef r0 As Single)

  beta = Radians(beta)
  gamma = Radians(gamma)
  Dim rm As Single = r0 - ((R1 - R) * (Sin(1)))
  Dim x1 As Single = (-R * Cos(beta)) + ((R1-R) * Sin(1) * Tan(beta))
  Dim x2 As Single = (R - R1) + (R1 * Cos(1))
  Dim y1 As Single = r0
  Dim y2 As Single = ((R1-R) * Sin(1)) + (R * Sin(gamma))
  Dim y3 As Single = ((R1-R) * Sin(1)) - R

  Dim Collision As Boolean = False

'ball-shaft collision
  If y3 < 0.01 Then Collision = True

'ball spacing collision
  If (((2 * (rm ^ 2)) ^ 0.5) - (2 * R)) < 0.01 Then Collision = True

'beta-gamma collision
  If gamma <= beta Then Collision = True

'x1-x2 collision
  If x1 >= x2 Then Collision = True

'Radius collision
  If R1 <= (R * 1.05) Then Collision = True

'y1-y2 Collision
  If y1 >= y2 Then Collision = True

  Return Collision
End Function
```

A.6 Quartic Derivation of Engine Torque Function

A basic quartic equation takes the form:

$$y = ax^4 + bx^3 + cx^2 + dx + e$$

Assuming that at zero engine speed (x), there would be zero torque output (y), the last term can be ignored. The 'd' term can be used to allow the shape of the curve to be controlled manually by the user and hence is known. An additional 3 properties of the curve are generally known:

1. The maximum torque output (x_1) is known together with the engine speed it occurs at (y_1)
2. The torque output (x_2) at maximum power output is known, together with the engine speed it occurs at (y_2)
3. The turning point of the curve (where $\frac{dy}{dx} = 0$) must occur at the maximum torque output (x_1), by definition.

The three known equations are thus:

$$y_1 = ax_1^4 + bx_1^3 + cx_1^2 + dx_1 \quad (1)$$

$$y_2 = ax_2^4 + bx_2^3 + cx_2^2 + dx_2 \quad (2)$$

$$\frac{dy}{dx} = 0 = 4ax_1^3 + 3bx_1^2 + 2cx_1 + d \quad (3)$$

From Equation (3):

$$2cx_1 = -4ax_1^3 - 3bx_1^2 - d$$

Hence:

$$cx_1^2 = -2ax_1^4 - \tfrac{3}{2}bx_1^3 - \tfrac{x_1}{2}d \quad (4)$$

Multiplying by (x_2^2 / x_1^2):

$$cx_2^2 = -2ax_1^2x_2^2 - \tfrac{3}{2}bx_1^1x_2^2 - \tfrac{x_2^2}{2x_1}d \quad (5)$$

Combining Equations (1) and (4):

$$y_1 = ax_1^4 + bx_1^3 - 2ax_1^4 - \tfrac{3}{2}bx_1^3 - \tfrac{x_1}{2}d + dx_1$$

Hence:

$$b = -2x_1a + x_1^{-2}d - 2x_1^{-3}y_1 \quad (6)$$

Combining Equations (5) and (2):

$$y_2 = ax_2^4 + bx_2^3 - 2ax_1^2x_2^2 - \tfrac{3}{2}bx_1x_2^2 - \tfrac{x_2^2}{2x_1}d + dx_2$$

Gathering terms:

$$y_2 = (x_2^4 - 2x_1^2x_2^2)a + (x_2^3 - \tfrac{3}{2}x_1x_2^2)b + (x_2 - \tfrac{1}{2}x_1^{-1}x_2^2)d \quad (7)$$

Combining Equations (6) and (7) and simplifying:

$$y_2 = (x_2^4 - 2x_1^2x_2^2)a + (3x_1^2x_2^2 - 2x_1x_2^3)a + (x_1^{-2}x_2^3 - \tfrac{3}{2}x_1^{-1}x_2^2)d$$
$$\ldots + (3x_1^{-2}x_2^2 - 2x_1^{-3}x_2^3)y_1 + (x_2 - \tfrac{1}{2}x_1^{-1}x_2^2)d$$

Gathering terms:

$$y_2 = (x_2^4 - 2x_1^2x_2^2 + 3x_1^2x_2^2 - 2x_1x_2^3)a + (x_2 + x_1^{-2}x_2^3 - 2x_1^{-1}x_2^2)d + (3x_1^{-2}x_2^2 - 2x_1^{-3}x_2^3)y_1$$

Hence:

$$a = \frac{y_2 - (x_2 + x_1^{-2}x_2^3 - 2x_1^{-1}x_2^2)d - (3x_1^{-2}x_2^2 - 2x_1^{-3}x_2^3)y_1}{(x_2^4 - 2x_1^2x_2^2 + 3x_1^2x_2^2 - 2x_1x_2^3)}$$
$$b = -2x_1a + x_1^{-2}d - 2x_1^{-3}y_1$$
$$c = -2ax_1^2 - \tfrac{3}{2}bx_1 - \tfrac{1}{2}dx_1^{-1}$$