Application of Vehicle-to-Infrastructure Networks in Vehicle Control and Monitoring Systems

Thesis by:
Szilárd Aradi

Budapest University of Technology and Economics
Faculty of Transportation Engineering and Vehicle Engineering
Department of Control Transportation and Vehicle Systems

Supervisor:
Prof. Péter Gáspár, DSc
Systems and Control Laboratory
Computer and Automation Research Institute
Hungarian Academy of Sciences

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

April 29, 2015
Acknowledgements

First of all, I would like to express my deepest appreciation and thanks to my mentor Károly Gynes who introduced me to the engineering profession and teaching.

I would like to express my deepest gratitude to my supervisor Prof. Péter Gáspár who showed me the way to the sophisticated and accurate research work and made a lot of valuable suggestions to improve my thesis.

Deepest gratitude are also due to Prof. József Bokor who ensured the opportunity for pursuing my doctoral studies and researches in his scientific school, as well as for his support and confidence during my departmental work.

I would like to thank Prof. László Palkovics who helped me to gain industrial experiences and supported my scholarship.

I highly appreciate the help of Tamás Bécsi who showed me the way to solve several difficult theoretical issues and he is tolerating for years that I am his officemate.

I am thankful to all my colleagues at the Department of Control for Transportation and Vehicle Systems and MTA SZTAKI.

I would like to express my major gratitude for Erika and for my family for their love and support. And last but not least I am extremely grateful to my Mother who – to my greatest regret – can no longer read this.
# Contents

## 1 Introduction
1.1 Motivation ................................................. 1
1.2 Overview of the Thesis .................................... 2

## 2 Server Solutions for On-line Telemonitoring Systems
2.1 Introduction ............................................... 5
2.1.1 Railway Application ................................... 7
2.2 Positioning the Telematics System in Enterprise Environment ........ 8
2.3 System Architecture ........................................ 9
2.3.1 Communication Model ................................... 11
2.4 The control of the telemonitoring systems’ server model ............ 13
2.4.1 Modelling ............................................... 14
2.4.2 Restrictions and simplifications .......................... 18
2.4.3 Control goals ............................................ 18
2.4.4 Controllability, observability ............................ 19
2.4.5 Control design ........................................... 20
2.4.6 Robustness ............................................... 24
2.4.7 Validation and simulation results ........................ 26
2.5 New results: Thesis I ........................................ 29

## 3 Development of Vehicle On-board Communication System for Rough Environment
3.1 Introduction .................................................. 30
3.2 Vehicular environment ....................................... 31
3.3 Transmission line principle .................................. 33
3.3.1 Reflections in transmission lines ......................... 34
3.4 Differential signalling technologies .......................... 35
3.4.1 TIA/EIA RS-485-A ....................................... 36
3.4.2 Controller Area Network ................................ 37
3.5 Laboratory Experiments ..................................... 38
3.6 Network look-up algorithm .................................. 40
3.7 Experimental results ........................................... 45
3.8 New results: Thesis II ........................................ 47

4 Design of Predictive Optimization Method for Energy-Efficient Operation of Trains 48
4.1 Introduction ................................................... 48
4.2 Motivation .................................................... 49
4.3 Energy Saving Possibilities .................................. 50
  4.3.1 General Analysis ......................................... 50
  4.3.2 Analysis of Driving Style ............................... 52
  4.3.3 Section Analysis ......................................... 53
4.4 Energy Consumption of Rail Vehicles ..................... 56
  4.4.1 Problem description ..................................... 56
  4.4.2 Mathematical formulation .............................. 57
4.5 Parameter Estimation ........................................ 59
  4.5.1 Proposed method ........................................ 60
  4.5.2 Error Analysis .......................................... 66
4.6 Development of the Optimization Algorithm ............ 70
  4.6.1 Simple reference ......................................... 70
  4.6.2 Reference considering track inclinations ............ 75
4.7 Driver Advisory Algorithm .................................. 84
  4.7.1 Model Predictive Optimization Problem ............... 84
  4.7.2 Optimization algorithm ................................ 87
  4.7.3 Remarks ................................................ 88
4.8 Simulation results .......................................... 88
4.9 Sensitivity and Robustness ................................. 90
  4.9.1 Robustness to Parameter Uncertainty ................. 90
  4.9.2 Optimization Weights .................................. 92
4.10 New results .................................................. 93
  4.10.1 Thesis III .............................................. 93
  4.10.2 Thesis IV ............................................... 94
  4.10.3 Thesis V ............................................... 95

5 Conclusions and Future Research 96
5.1 Practical Aspects ........................................... 96
5.2 Future research ............................................. 97

List of Figures .................................................. 101
List of Tables .................................................. 103
References ...................................................... 104
Chapter 1

Introduction

1.1 Motivation

Demand for transportation is continuously increasing in the European Union resulting in an increase in the energy consumption of the sector. The global challenge is to make transportation more efficient and environmentally friendly. Europe is accelerating its progress towards a low-carbon society [European Commission, 2011] in order to reach the target of an 80% reduction in emissions below the 1990 levels by 2050. To be able to achieve that, emissions should be reduced by 40% by 2030 and 60% by 2040. Each sector must contribute to achieve this, and transportation has an essential role in this effort. Another important aspect is that the major part of the transport demand growth affects road transport. It results in a strong competition between rail and road transportation, in whose effectiveness, operation costs and reliability are important factors.

Owing to the rapid development of microelectronics and mobile telecommunications at the end of the '90s a wider range of fleet diagnostics and satellite tracking have become available. These innovations provided a technological background for the creation of on-line telemonitoring systems. Economic demand for telemonitoring systems rose as a consequence of the increased competition in passenger and freight transport especially in road traffic. This effect was reinforced by the growth in traffic density and demand for transportation in Europe. One of the social impacts of this progress is an increase in the number of accidents, which increases the demand for safer vehicles. The spread in on-line telemonitoring systems was greatly aided by the constant decrease in communication charges and the increase in the speed of mobile data transmission. In addition to the traditional advantages (e.g. door-to-door service) these systems have increased the competitiveness of road transport with shorter delivery time, less delay, better cargo tracking and more cost-effective operation.

An on-line telemonitoring system integrated with a modern ERP system (which naturally considers the special requirements of the rail transport) could also improve the efficiency of a rail traction company. In addition the savings can be further increased with the introduction of the electronic documents. It is particularly true in the rail operation, due to the large amount of documents required for safety reasons. As it can
be seen a complex IT system is needed, which can cover the classic ERP functions, the special requirements of the railway companies and the collection of data concerning the vehicles.

Besides the general efficiency the energy saving possibilities should be emphasized. Several possibilities are available for construction companies and railway operators, such as design considerations and advanced train-control and intervention systems. The infrastructure operators can cut costs in several ways. The main consumers are the equipment of the stations such as lighting, train and engine sheds, switch point heating etc. Another option is to use modern timetable design methodologies to avoid the unnecessary decelerations and stops, but it only works if the engine driver can hold the timetable precisely. A further important aspect in the rail traffic is the utilization of the inclinations and the optimal selection of the train speed considering the speed restrictions looking ahead to the whole section.

The introduction of such systems and solutions requires significant investment. It is important to note that the Hungarian State Railways are in a special situation compared to the Western European railway companies. Financial resources are scarce, so the infrastructure and rolling stock are becoming obsolete. The ongoing maintenance of these old instruments has high costs, since the safety and operation of the rail must be maintained. Other cost-increasing factor is the higher tractive energy consumption caused by the even denser speed limits due to the deterioration of the infrastructure. In this situation increasing the efficiency and reducing the costs are essential to provide a basis for at least a portion of the required investment funds.

Considering all the problems of railway systems would lead to a very complex and diverse task, however it can be partitioned into smaller subtasks or special areas that can be handled well.

I used a vehicle-side approach addressing the vehicle as a source of data and started to examine the solutions in the field of road transport. The aim was the development of a complete remote monitoring system intended to deal with the architectural issues, communication technologies and server-side load issues. After the design of the system it was important to examine the usability of the data for the derived high-level functions. This area is also very sparse, as virtually the entire railway operation, as well as accounting, marketing and controlling functions can be supported. From these possibilities I chose the individual vehicle control issues as some of the most relevant issues of train operation, only after the safety aspects. The Hungarian State Railways do not use any driver advisory system, neither do they encourage economical driving, or deal with this topic in the training of the engine-drivers. This is the reason why I addressed this topic as my research interest.

1.2 Overview of the Thesis

The rapid development of the road transport telematics has also affected the railways, as it can offer similar benefits, however it can meet different requirements in several aspects. During my research I dealt with road and railway information systems. The methods
presented in my thesis can solve special problems in the railway remote monitoring systems and offer several options for further development.

The first part deals with the application possibilities of remote monitoring systems for railway examining the benefits of the introduction of such a system. Through the example of the Hungarian State Railways I present the positioning of the monitoring system within the railway IT system. Communication with the vehicles must meet several quality requirements including availability, response time or safe recovery after a possible shutdown, etc.; which imposes strict requirements on the system’s server application. Thus in addition to careful planning - for which a possible structure is introduced - the control of the service is needed, which can be handled using control theory [Abdelzaher et al., 2003],[Abdelzaher et al., 2004], naturally at the presence of an appropriate queuing model [Qin et al., 2009]. In IT systems most of the state variables can be measured directly, although in many cases system load requires estimation [Robertsson et al., 2004]. To solve this problem I modelled the system as a M/G/1 type queuing model and designed a state-feedback control system).

Locomotive on-board computers with network access can serve not only the communication needs of the engine and the driver, but can also act as a gateway to serve the entire train telemetry subnet and support - including but not limited to - the maintenance or even passenger information [Edwards et al., 2005] tasks. However, train composition is not constant, and - with the exception of the newer types of carriages - free-to-use wires inside the remote control cables are not provided. The next part of the thesis deals with wired networks issues of the trains without the so-called 'intra-train' network in special circumstances and harsh environments [Russo et al., 1993]. However the implementation of the digital data transmission on a non-standard communication is an important problem to be solved to maintain the appropriate levels of the network quality parameters. First the non-standard solutions were tested under laboratory conditions. In addition I proposed an appropriate sectioning solution to detect the variable network topology and I designed an algorithm to control the detection process. This will ensure the communication possibility via the free wires of the train remote control cable. The developed methods were validated on real trains both in stationary and moving states.

The availability of on-line and real-time information opens a huge opportunity for increasing the efficiency of the railway operation. The optimization of the operation of the railway system can reduce delays caused by the unique chain-reaction-like delay propagation and helps to restore the normal operation. In addition it can help the energy optimization of the individual train movements [Howlett, 1990], [Howlett et al., 1994], [Liu and Golovitcher, 2003].

The last part of my thesis focuses on the examination of the traction energy consumption between two stations. To achieve this the real-time telemetry data provided by the monitoring system are essential, so the available journey time can be determined, but it is also necessary to know the speed limits and the slope conditions of the relevant section. The task includes two conflicting criteria, which keep the journey time and minimize energy consumption. I have developed a solution to apply predictive optimization, which has induced some further problems to be solved.
The first task was the determination of the propulsion resistance of longitudinal train dynamics. I applied the polynomial function of the resistance force found in the relevant literature [Hay, 1961], [Iwnicki, 2006], [Rochard and F., 2000]. To determine the parameters I have used on-board engine data with one-second sampling time and proved that the parameters can be well estimated without additional special measurements.

To ensure the time-keeping property of the predictive optimization it is essential to generate a reference run [Grüne and Pannek, 2011], which keeps the required journey time, thus the optimization algorithm can follow it on the prediction horizon as a constraint. For this purpose a formerly recorded train run can be used or it can be generated with a suitable algorithm [Howlett et al., 2009] [Howlett and Pudney, 1995]. I developed two optimal speed profile generator algorithms which consider the entire section and the journey time. The first one is simpler with less processing requirements, but it ignores the inclinations of the track, while the second method extends it with the slopes at the cost of higher computation demand.

The final goal was to develop a driver advisory system which can support energy optimal driving by providing real-time recommendation to the engine-driver. Fundamentally a pre-calculated, static speed profile could be essential, but it can not be actuated by the driver in all cases, thus the optimal profile may become obsolete. During the journey the continuous optimization is expedient which can be achieved by the proposed predictive control system.

Finally the some practical applications of the results are outlined and future research directions and tasks are given.
Chapter 2

Server Solutions for On-line Telemonitoring Systems

2.1 Introduction

An on-line telemonitoring system (often called Fleet Management System - FMS in the field of road transport) collects, stores and provides complete comprehensive information about the current state of vehicles and cargo, route history, expected events, as well as driver activities for vehicle maintenance and operator companies. The main fields of application are the following:

- Vehicle operation
- Traffic management
- Traffic safety
- Security of freight
- Environmental protection.

As mentioned in the introduction the spread of on-line telemonitoring systems was greatly aided by the constant decrease in communication charges and the increase in achievable bandwidth in mobile data transmission. For all these reasons after the 2000s on-line telemonitoring systems became widespread rapidly. Their advantages are the following:

- A greater safety in delivery
- Aiding dynamic freight arrangement
- Constant tracking of the mechanical condition of the vehicles
- Reduction of operational costs (fuel consumption and maintenance costs)
• Avoidance of the illegal usage of the vehicle and the fuel manipulation
• Easier documentation (e.g. journey log)
• Driver motivating system (driving style analysis)
• Developing safety of traffic (speeding and accident detection)
• Increased security of freight
• Better environment protection

Naturally the main reason for the installation of an on-line telemonitoring system is the expected decrease in the overall operating costs of the company by improving efficiency. Although the establishment of such a system requires one-time costs of investment and its operation also imply costs, these expenses are counterbalanced by the savings. The cost reductions arising from the more efficient operation are typically the following:

• Reduction of operation costs thanks to the increase in vehicle utilization, optimal route planning and minimal journey time considering the service time of the drivers.

• The on-line telemonitoring system provides the possibility of the maximal utilization of the vehicle parts’ lifetime by the complex monitoring of the vehicle. Simultaneously it warns for the necessity of the replacement and supports the logistics solutions also.

• The system may improve the quality of the estimation of the fuel norms through the measurement of the real consumption.

The possibility of on-line vehicle tracking is an ideal tool for the optimization of the logistic processes. Not only the route of the goods can be determined, but the vehicle movements can be optimized. The improvement in logistics costs could be the followings:

• Improve efficiency of freight transport
• Decrease of extra charges caused by the delays.
• Significant decrease in vehicle stop time can be reachable resulting in cost reduction with keeping the same transportation performance.
• Monitoring helps forcing back the illegal utilization of the company’s vehicle.

Another requirement is to provide the safety of the freight through using on-line telemonitoring system. The data acquired make possible to follow or retrace the vehicle events. With using on-line system it is possible to intervene in critical situations and to monitor the keeping the traffic regulations. The following section presents the
kind of functionality and the system architecture that can satisfy the above mentioned requirements.

Initially these systems were introduced by the international transport companies with heavy commercial vehicles (large goods vehicle). These early systems have collected only GPS positions and sent them by SMS to a central server. With the reduction of costs the FMS solutions have broken into the segment of the smaller commercial vehicles (light commercial vehicles). Nowadays the FMS is used almost in all vehicle segments, typically in the following ones:

- Large goods vehicle (commercial vehicles over 3.5 tons)
- Light commercial vehicles (commercial vehicles not more than 3.5 tons)
- Passenger vehicles
- Construction and agricultural machinery

2.1.1 Railway Application

The use of telemonitoring systems is becoming widespread in the area of road transport, and almost all road transportation companies use some kind of real-time fleet management system operated either by itself or a third-party provider. The need is also arising in the field of rail transportation, and it also affects the Hungarian State Railways (MÁV) also. MÁV have developed several systems in the past few decades for keeping pace in the race for competitiveness in the area of transportation.

The problem emerges as the new informatics systems of the railways need more and more precise and up-to-date information of the fleet. The handling of the engines was managed by the Traction Subsidiary Company of the State Railways in the mid 2000s, and its previous logbook system handling all information of the trains was ’paper-based’ meaning all information was generated on paper forms and was entered into the informatics system after the service of the driver or engine ended. Naturally these data were nor on-line, nor real-time. Naturally this system did not meet the requirements to interface with other systems, such as the infrastructure management, order registering or maintenance systems.

The development and implementation of the on-line telemonitoring system at the Hungarian State Railways started in 2006 with a pilot project under the cooperation between the MÁV and the Department of Control and Transportation of the Budapest University of Technology and Economics. The experiences of this project led to the development of the final system for the MÁV which was started in 2007. Nowadays all engines are equipped with an on-board unit, registering the position of the train and measuring the main engine and cockpit attributes and also the operations of the engine drivers.

First, the roles of and the positioning of such system is presented amongst the informatics systems of a rail company through the example of the Hungarian State Railways. This is followed by the modeling and design aspects of a Fleet Management system for
rail transportation needs with special regard given to the communication server’s design. In the second half of the chapter a queue model based server model control design is presented for the purpose of ensuring flawless response of the communication server under computational resource restrictions.

2.2 Positioning the Telematics System in Enterprise Environment

The telemonitoring system takes its place in the informatics system of the Hungarian State Railways as can be seen in Fig. 2.1. The telemonitoring system is part of the IT system of the Traction Company, since this firm possesses and manages the engine fleet of the Hungarian State Railways.

![Diagram of the telematics system positioning](image)

Figure 2.1: Positioning the telemonitoring system in the global informatics structure of MAV

The system is part of the Electronic Logbook System (ELS), which manages all communication with the fleet and stores all data received from the trains.

The ELS is connected to several systems (though for simplicity reasons, all connections are not marked on the figure), but it main interface leads to the Traction Technology Planning System (TTPS).

It can be said, that the TTPS is the main core system for the Traction Co., since the TTPS handles the engine drivers’ service and working time planning, the planning of the engine turns and allocation. Therefore the TTPS must be connected with the SAP system of the Railway Co., the Order Registering and Management System (ORMS) and also with the Maintenance and Repair Systems of the Maintenance subsidiary company of the Rail Co.
The ORMS handles the ordering procedure. Its main task is to receive the main attributes of the orders, and after transforming it to preliminary schedule, and transfers it to the TTPS (and to the ELS) systems. Naturally it is connected to the SAP also, and to the main Customer Companies, such as the “MÁV Start”, which is handling the passenger, and “MÁV Cargo” which is handling the freight transport services.

(Note: The status and the proprietary of the MÁV companies has been changed several times in the recent years. Now the MÁV Cargo operates as a part of Rail Cargo Group and called Rail Cargo Hungaria. The Traction Subsidiary Company, MÁV Trakció has been fused to MÁV Start.)

### 2.3 System Architecture

The construction of the on-line telemonitoring system [Aradi, 2007] is demonstrated in Fig. 2.2. The two main components are:

- On-board units,
- Central server.

The operation of the system is the following. The on-board units (OBU) on the locomotive measure the operational parameters of the locomotive (state of the switches, energy consumption, motor parameters, etc.), and its position (aided by GPS based location), and they store the data given by the engine-driver (the name of the actual activity, etc.). These parameters are sent to a central server at the actualization of previ-
ously defined events (alarm-signal, sudden decrease in fuel level, etc.) and in previously defined periods of time.

![Image of MMI of on-board unit](image)

**Figure 2.3: The MMI of the on-board unit**

On-board computers communicate with the central server through GSM network. The incoming data are evaluated and stored in a database. If necessary the central server can send an alarm to a given e-mail address or even a mobile phone. In this structure communication from the server towards the locomotive is plausible as well. Aided by this the incoming data packages can be confirmed, a written message can be sent to the engine-driver and the parameters of the board unit can be set.

Locomotives are detectable and observable almost constantly (on-line) and the operating parameters (running performance of vehicles, energy consumption, activities and work time of drivers, delivery performance) can be followed by a later evaluation of data stored in the centre (offline).

The most important aspects of the building of an on-board unit are heavy-duty design (EMC protection, shake protection, fluctuation of environmental temperature, etc.) and modularity. Therefore one should use a system that is built up of individual units. The connection of these by a series communication connection is worth realizing for the sake of simplicity and easy expansion. For this the most appropriate is the Controller Area Network (CAN) bus system.

Board unit [Marczinák and Dicső, 2006] is made up of the following main units:

- GSM/GPS module,
• Central unit,
• Incoming unit,
• Human interface device,
• Diagnostic adapter,
• CAN bus,
• Power supply unit and background batteries.

The Man-Machine Interface of the OBU is shown on Fig.2.3. The OBU was designed and manufactured for MÁV Trakció by Prolan Co., a Hungarian manufacturer.

2.3.1 Communication Model

The communication system can be built up by the OSI model [ISO, 1994] as shown in Table 2.1. The connection point between the OBU-s and the server is the session layer (TCP socket).

Essentially the network model is built on IP based technologies and protocols. The OBU communicates with help of the GSM based internet service. The modems use General Packet Radio Service (GPRS).

Considering the safety aspects [Bécsi and Aradi, 2008] the OBU-s connect to a dedicated GSM APN and there is Virtual Private Network (VPN) between the internet service provider (ISP) and the central server.

In our model the Transmission Control Protocol (TCP) is used in the transport and session layers because of the following advantages in relation to the User Datagram Protocol (UDP):

• Ordered data transfer,
• Retransmission of lost packets,
• Discarding duplicate packets,
• Error-free data transfer,
• Congestion/Flow control.

In the data block of TCP packet a record structure was built up, which contains the data of the locomotive and the train. For the declaration of the structure and data types a standard XML schema [W3C, 2012a] was created.

For the reliable communication an XML [W3C, 2012b] based protocol was designed with the following main properties:

• The OBU has to initiate the communication. (The OBU-s are the clients, the center is the server.)
Table 2.1: The 7-layer model of the communication subsystem

<table>
<thead>
<tr>
<th>OSI model</th>
<th>Used protocol or service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical layer</td>
<td>GSM, 100BASE-TX</td>
</tr>
<tr>
<td>Data link layer</td>
<td>GPRS, Ethernet</td>
</tr>
<tr>
<td>Network layer</td>
<td>Internet Protocol (IP), Virtual Private Network (VPN)</td>
</tr>
<tr>
<td>Transport layer</td>
<td>Transmission Control Protocol (TCP)</td>
</tr>
<tr>
<td>Session layer</td>
<td>TCP socket</td>
</tr>
<tr>
<td>Presentation layer</td>
<td>UTF-8</td>
</tr>
<tr>
<td>Application layer</td>
<td>XML based protocol</td>
</tr>
</tbody>
</table>

- The link is permanent; the disconnection is permitted only if an error occurs.
- If there is not any working link, it needs to be established as soon as possible.
- The protocol uses the XML 1.1 standard. The data must be validated by an XML Schema Definition (XSD) file. Only the convenient data can be accepted.
- Every message must be acknowledged. The sending of a message has to be repeated until acknowledgement. (The timeout is 30 seconds.) The errors caused by frozen TCP sockets can be eliminated by this method.
- Every message must contain an ordinal number. The persistence of this number needs to be checked for finding absent messages.
- The protocol is change-oriented. Only the changed data should send to the server. This technique can decrease the communication costs.
- Every message must contain the following data:
  - Ordinal number
  - Timestamp (UTC)
  - GPS data (validity, longitude, latitude, heading, speed)
- At the start of the communication the OBU must login to the server with its own individual identifier.

2.3.1.1 Functionality

The telemonitoring system, as its first version has the following functionalities:

- Self identification, where the OBU identifies the engine UIC (International Union of Railways) number to the server.
- Data sending, where the OBU sends the collected data to the server.
- Base data group: message UID, timestamp, GPS position, speed and heading.
- Technical parameters data group: engine speed, engine temperature, energy consumption, switch states, etc.
- Activity data group: train data, customer data (any service related data, if not given by server).

- Message from the server to the driver, where the operator sends non-formal message to the driver.
- Message from the driver to the operator, same as above, opposite way.
- Service start-stop messages, Service start/stop notifications for engine drivers and other workers (including human authentication and authorization).
- Data request from the OBU, where the OBU requests for data stored on the central system.
  - Accumulated Fuel level, if there are more engines connected together into traction formation.
  - Train Data, train number, accumulated weight, etc.
- Emergency alerts: Both generated by the OBU or the engine driver. (fire, accident, malfunction, detected anomalies, etc.)
- Geofencing: where the OBU detects the entering/leaving of a previously given area, and sends notification. (stations, point areas, crossings, etc.)

### 2.4 The control of the telemonitoring systems’ server model

The current section deals with the control of the communication server of a telemonitoring system introduced in the previous sections [Aradi and Bécsi, 2008], [Bécsi and Aradi, 2009] with special regards to performance, availability and quality of service.

Systems of this kind are generally handled as the combination of an M/G/1 queuing [Abdelzaher et al., 2004][Hellerstein et al., 2002] and a load model. The queuing model is responsible for the modelling the length of the queue between the incoming internal and external requests, and their service provider i.e. it determines how many unserved tasks the software realizing the system still has (see Figure 2.4).

The load model, on the other hand, determines how large the amount of the system’s resources has to be in order to be able to execute these tasks [Abdelzaher et al., 2003][Robertsson et al., 2004]. In case of a server application, the load means the utilization of hardware-based resources, which can be processor load, memory usage, I/O operations of the computer storage, and in a given case, use of network resources.

The question should arise that what is the motivation of controlling a simple server queue where the system has no impact on the arrival of request. Though, in a real system of finite resources the service of incoming requests is only one task a server must
handle therefore it has to regulate itself to keep resources to other processes. These two contradictory tasks can be summarized as:

1. The fundamental role of the communication server is to provide service to incoming requests as fast as possible.
2. It has to prevent the overload of the server resources by restricting the number of served request in a cycle, allowing enough computational capacity for other subtasks of the server.

The task therefore shows similarities to highway ramp metering, or network congestion management of Ethernet routers where the server has to restrict the input traffic to ensure the congestion free flow of the interior system.

In order to control service performance, the expected processor load has to be estimated during the construction of the system through the estimation of service time per request, meaning that during the realization, a classical measuring problem has to be realized in which the number of the server’s clients, the cardinality of messages sent and to be received by them, and the resource need of these services has to be determined as accurately as possible.

Extreme over-sizing of the resources should handle these problems however it has investment and maintenance cost. Though system expansion could affect even a well-scaled structure leading to a similar situation:

- Increasing client number;
- Growth in request rate per client because of service extensions;
- The rise of service time requirement because of the improved tasks in additional system of sharing the same resource.

Naturally after a careful resource planning the system should serve every request as fast as possible. However, should any incident of vis major nature occur e.g. an intended shutdown, the system had to face such a large number of requests that their execution could result in an overload, or even the breakdown of the system.

A telemonitoring system without congestion control, cannot respond to a challenge of this sort, so the installation of a service model is necessary which can take control of the load of the resources resulting in a longer service time yet improved availability ensuring the stable operation of the system.

2.4.1 Modelling

The first step of control design is the generation of the proper mathematical model of the system. The internal dynamics of the system have to be mapped identifying the inputs, state variables and the control goals. According to Figure 2.4 the simplified system can be represented as a G/G/1 queue model, a simple server queue where the arrival and the service intervals are having some kind of general or unknown distribution. Having
relatively high number of clients with both periodical and event based requests from independent peers the arrival process should be considered as a Poisson process leading to the queue model M/G/1. In an idealistic environment the service times for each request should be considered as constant leading to M/D/1 queue model, though in the real application environment this assumption can not stand. Generally an independent server solution would operate with deterministic service rates with only minor deviations though in case of server overload, serve intervals could grow significantly. This feature indicates that in the preliminary mathematical formulation the service times could not be modelled as a simple parameter with minor disturbance, it is much more convenient to include it as a system state that is unfortunately but inherently not controllable.

The server system operates in service cycles, meaning that though the arrival of the requests is continuous and is handled by a separate thread, their processing is made in cycles, where the number of requests that are going to be handled must be determined in advance ie the requests are served in batches.

Figure 2.4: M/G/1 queue model of the telemonitoring service with one server

To find the adequate mathematical formulation and modelling technique the main characteristics and behaviour of the system should be examined.

Though the requests are served in batch groups the superposition principle stands ie the service time is proportional to the number of served request therefore linear approach will be applied.

The system could be handled in continuous time domain since the arrival mode of the requests; though it is much more convenient to focus on the service process which is operating in cycles. This feature indicates that the system model should be discrete time where each step corresponds to an application cycle.

The system’s response time highly depends on the load of the general system that includes the server application, therefore some parameter uncertainty should occur in the system’s model. Though this phenomenon has slow dynamics and so the system will be treated as deterministic and the change of the parameters affecting the behaviour will be handled in the control design phase. Because of the uncertainty in the parameters of the service times time invariance can not be strictly stated though the basic behaviour is not affected by the changes only the speed of the system dynamics, therefore in modelling phase the system will be treated as time invariant.

It seems to be effective to suppose that the system is dynamic, time-invariant, de-
terministic and discrete time. The system can of course lose some features while being constructed but it is definitely practical to formulate these presuppositions.

During the construction of the model, every parameter has to be set which together can influence the operation of the system. It is important to know from the point of view of the resources how many clients the system serves at the same time. The client number \( x_{\text{cln}} \) is a simple integrating state variable whose value in step \( (n+1) \) depends on the client number of the previous cycle \( (n) \), and on the number of incoming and outgoing clients (\( d_{\text{cli}} \) incoming clients; \( d_{\text{clo}} \) outgoing clients) during the cycle:

\[
x_{\text{cln}}(n+1) = x_{\text{cln}}(n) - d_{\text{clo}}(n) + d_{\text{cli}}(n)
\]

(2.1)

The amount of the incoming requests in a cycle - be they requests from the side of the clients, or originating at the server - depend on the actual client number, yet it is still practical to define these as external inputs as the density of occurrence of these requests can vary in time. The queuing model of the requests waiting for service in the given cycle can be written up from this \( x_{\text{ql}} \), which can be written up from the value taken up by it in the previous cycle, and the incoming \( (d_{ri}) \) and outgoing/served \( (x_{\text{serve}}) \) requests:

\[
x_{\text{ql}}(n+1) = x_{\text{ql}}(n) - x_{\text{serve}}(n) + d_{ri}(n)
\]

(2.2)

The number of requests which are to be served in the next cycle are interpreted in the given cycle as the system’s input i.e. the state variable depends on the number of requests which are to be served \( (u_{\text{serve}}) \):

\[
x_{\text{serve}}(n+1) = u_{\text{serve}}(n)
\]

(2.3)

In order for all the information to be available, the state variable of the resource load has to be determined as well. What the resources regards, the demand for system memory can be regarded as an insignificant variable because this resource, as it is also shown by the examination of the sample system provided with the suitable data structure, becomes practically inexhaustible. The network and the computer storage I/O operations, as the CPU is waiting for the realization of these, can be converted unambiguously into CPU load values. Every universal built-in algorithm i.e. searches, maintenance of lists, control of state machines, etc. occupy the CPU’s time as well. According to these considerations, it is suitable to model the resource need with CPU time or with CPU load [%]. This value depends on more parameters:

- The load independent from other variables \( (t_{\text{sb}} \text{ Stand by [%]}) \) This load summarizes all activities of the server that are independent from the queue though essential for the service process such as self maintenance, the management of supplementary search lists network handling and logging, etc.;
• The universal functions applied to certain clients, which is a function of the client number \( x_{cln} \) multiplying \( c_{ekp} \) passive clients [%/client]. This is a necessary addition for the load, since the server must pay attention to the connected clients even when they do not post requests;

• The CPU load created during the login and logout of the clients \( (d_{clo} + d_{cli}) \) multiplying \( c_{cico} \) clients in/out load [%/client]). The connection handling could also be treated as a special request operation of the queue though because of its different nature from the regular service it is handled separately;

• The load that occurs at the receiving of requests. This is a basically the load of enqueuing. \( (d_{ri} \) multiplying \( c_{ib} \) request receiving load [%/client] )

• And finally the load of requests served in the given cycle \( (x_{serve} \) multiplying \( c_{serve} \) service load [%/client]) This load consist of the main request processing tasks, such as syntactic and semantic interpretation, request audit, response creation and handling etc.:

\[
x_{proc}(n + 1) = t_{sh} + c_{ekp}x_{cln}(n) + c_{serve}x_{serve}(n) + c_{cico}(d_{clo}(n) + d_{cli}(n)) + c_{ib}d_{ri}(n)(2.4)
\]

This last state variable is also taken as the model’s output i.e.:

\[
y(n) = x_{proc}(n) \tag{2.5}
\]

The equations 2.1-2.5 define a linear system with the state space representation of these Linear equations is as follows:

\[
x(n + 1) = Ax(n) + Bu(n) \tag{2.7}
\]

\[
y(n) = Cx(n)
\]

As can be seen from the state space representation of the model in equations (2.42) and (2.7) the server model is a discrete linear system with four states \( (x_{proc} \) the processor load, \( x_{serve} \) the served requests in the cycle, \( x_{ql} \) the queue length and \( x_{cln} \) the number
of connected clients) and four inputs ($u_{serve}$ the control input for the requests to be served, $d_{ri}$ the number of incoming requests and $d_{clo}, d_{cli}$ are the number of outgoing and incoming clients). The notations of input distinguishes $u_{serve}$ and the other three ($d_{ri}, d_{clo}, d_{cli}$) for reason, since the only input for the model that the control could have impact on is the number of the request to be served. The system has no influence on the other three inputs so they will be treated as disturbance. However their values are measurable and can be used in the control design process.

### 2.4.2 Restrictions and simplifications

Though the system is formulated as a linear state space, its value set has some restrictions. Also, for the design of a feasible control some simplifications are also needed. The value which is represented in the linear system as client number can under the effect of appropriate theoretical inputs take up negative values as well. Naturally this is not possible in a real system, and also could lead to negative CPU load as the result of system dynamics, which is also not realistic. Thus, client number has a minimum:

$$x_{real}^{cln} = \max(x_{theoretical}^{cln}, 0) \quad (2.8)$$

The same applies to queue length, which also cannot take up negative values:

$$x_{real}^{ql} = \max(x_{theoretical}^{ql}, 0) \quad (2.9)$$

Besides, the request number which is actually to be served in the given cycle does not only equal the theoretical service number i.e. that which is defined as input, but it is also limited by the number of serviceable entities in the queue:

$$x_{real}^{serve} = \min(x_{theoretical}^{serve}, x_{real}^{ql}) \quad (2.10)$$

The discrete time linear system which is taken up according to equation (2.42) has four inputs, $d_{do}, d_{cli}, d_{ri}$ and $u_{serve}$. However, it is only the last one of these, the service number, which can be used for controlling purposes, as the system has no impact upon the login or logout of the clients. Therefore, the effect of these inputs has to be interpreted as disturbance as been mentioned before.

In the system under investigation the significant part of the CPU load is mostly made up of the load of served requests. The service time of one request ($c_{serve}$) however depends on many external factors e.g. response time of the database and the occupation of other hardware resources. However, this system parameter thus cannot be considered to be constant, it is only its theoretical minimum that can be defined.

### 2.4.3 Control goals

The design of the system’s control has to be performed mainly by taking the above mentioned two basic aspects into considerations, so that the condition of keeping the positive value of input $u_{serve}$ has to be continuously satisfied.
The CPU load clearly determines the length of the server’s processing cycle. Should the system serve too many requests in a cycle, the time of the given cycle becomes too long, and, because of the service, the system cannot take care of other tasks e.g. login or communication of the clients, or the processing of requests. Thus, it is practical to determine a cycle time in which in the given cycle every task can be executed in the case of every parameter and input corresponding to regular operation. This is a simple measuring task, and the hardware need of the system should be determined according to this parameter. If the system is overloaded, the amount of requests that are to be served in one cycle needs to be determined so, that the CPU load should not be higher than the value of the determined cycle time. The same goes for the reverse case, should the CPU load be much lower than the cycle time, the system load is not optimal, the queue of pending requests does not decrease in the highest possible degree. However, exceedance of the cycle time is to a small extent and on a single occasion is allowed. Different surveys touch upon various fields of control theory while controlling systems of this kind, so both solutions of PI-type [Sha et al., 2002] and LPV-based ones can be found. [Qin and Wang, 2007][Qin et al., 2009]

2.4.4 Controllability, observability

The states and inputs of the system (2.42) are measurable, as it is an information technological application, so the application of an observer is not necessary. As the system can only be controlled with input $u_{serve}$, controllability analysis has to be performed on this input by using Kalman’s controllability rank condition:

$$B_s = [0, 1, 0, 0]^T$$

(2.11)

$$\text{rank}(C(A, B_s)) = \text{rank} \left( \begin{bmatrix} B_s & AB_s & A^2B_s & A^3B_s \end{bmatrix} \right) = 3,$$

where $B_s$ is the input selector of $u_{serve}$.

The rank of the controllability matrix that is projected to the system’s input $u_{serve}$ is three, which is less than the dimension of the system. As the system is not complete state controllable, the stability of the system cannot be ensured with input $u_{serve}$ under the present modelling paradigm. The unreachable state is the client number ($x_{cln}$), and therefore, the CPU load is not completely controllable either.

2.4.4.1 System reduction

As mentioned above, client number is not controllable, though measurable, and with previous calculations and measurements its effect on the CPU load can be estimated. Though it is a process that is out of the scope of the control design its calculated value can be used in the generation of the reference input of the control loop, or to take it for a disturbance which appears at the output of the system. Therefore the elimination of the client number is admissible. This can be easily kept in hand if we know the nature of the disturbance. In the present case, there is even something more that can be achieved, as we can determine this disturbing sign, which has an effect on the CPU load, with approximate estimating.
By eliminating disturbance inputs \((d_{clo}, d_{cli}, d_{ri})\) and the unreachable state \((x_{cln})\) the state space representation of the system is defined by the following matrices:

\[
\tilde{x}(n+1) = \tilde{A}\tilde{x}(n) + \tilde{B}\tilde{u}(n)
\]

\[
\tilde{y}(n) = \tilde{C}\tilde{x}(n),
\]

where \(x_{proc}, x_{serve}, x_{ql}\)

\[
\tilde{x} = \begin{pmatrix} x_{proc} \\ x_{serve} \\ x_{ql} \end{pmatrix}, \quad \tilde{u} = (u_{serve}), \quad \tilde{A} = \begin{pmatrix} 0 & c_{serve} & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}
\]

\[
\tilde{C} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}
\]

The system therefore becomes state controllable, as it satisfies the rank condition:

\[
\text{rank} \left( C(\tilde{A}, \tilde{B}) \right) = \text{rank} \left( [\tilde{B}|\tilde{A}\tilde{B}|\tilde{A}^2\tilde{B}] \right) = 3
\]

### 2.4.5 Control design

The rank of the (2.14) system’s controllability matrix is 3, which is equal to the dimension of the system, so the reduced system becomes controllable. The poles of the system, the eigenvalues of \(\tilde{A}\), are not stable as it can be seen on (2.15), so it is necessary to use state feedback in order to stabilize them.

\[
\overline{P}(\tilde{A}) = (0, 0, 1)
\]

As there are a solution to the pole-replacing task is given by Ackermann’s formula:

\[
K = \left( \begin{array}{ccc} 0 & 0 & 1 \end{array} \right) e^{-1}(\tilde{A}, \tilde{B})\phi_c(\tilde{A})
\]

(2.16)

For a given set of desired poles \((\alpha_1, \alpha_2, \alpha_3)\) state feedback gain \(K\) is determined as:

\[
K^T = \left( \frac{\alpha_1 \alpha_2 \alpha_3}{c_{serve}}, 1 - \alpha_1 - \alpha_3 - \alpha_2, -1 - \alpha_1 - \alpha_3 - \alpha_2 - \alpha_1 \alpha_3 - \alpha_1 \alpha_2 - \alpha_2 \alpha_3 - \alpha_1 \alpha_2 \alpha_3 \right)
\]

(2.17)

From practical considerations, the feedback of the state \(x_{serve}\) could be omitted. To eliminate the feedback of the state the second element of vector \(K\) should be 0. In this case, the following rule has to be observed while setting the poles:

\[
\alpha_2 = 1 - (\alpha_1 + \alpha_3)
\]

(2.18)

In this case, the value of \(K\) changes as follows:

\[
K = \left( -\frac{\alpha_1(-1+\alpha_1+\alpha_3)c_{serve}}{c_{serve}}, 0, 2\alpha_1\alpha_3 - \alpha_1 + \alpha_1^2 - \alpha_3 + \alpha_3^2 - \alpha_1^2\alpha_3 - \alpha_1\alpha_3^2 \right)
\]

(2.19)

As an example, by choosing an appropriate pole triplet that is stable and fast enough:

\[
\alpha_1 = \alpha_3 = 0.2; \alpha_2 = 0.6
\]

(2.20)

The vector of the state feedback will be as a function of service response time \(c_{serve}\) as follows:

\[
\tilde{K} = \begin{pmatrix} 0.024 \\ 0 \\ -0.256 \end{pmatrix}
\]

(2.21)
2.4.5.1 Keeping the cycle time

The structure mentioned in section 2.4.5 can adequately control the system into a stable state if it functions properly when - according to system scaling - less requests appear that the server constantly can serve. Naturally the contribution of this solution is minor in case the requests are arriving at the rate that had been previously planned and the service time of each request is also kept around its previously defined value.

However, should queue length become too high because of a rapid growth of request density or that of specific service time, the controller would “push” the queue length state variable over the 0 value, and so the cycle time of the system would grow as well. The lengthening of the cycle time can hinder the execution of the system’s other tasks and by doing this, it blocks proper functioning.

To prevent an overload of this sort, it is inevitable to use an alternative control loop, which puts emphasis on other aspects. The overload can not be permanent, as the increase in request rate or the decrease in server resources are temporary, and therefore the service gradually catches up with the appearance of requests, so the queue of pending requests decreases after the solution of the problem to normal value; it is practical to design a control which can ensure the availability of the system in the period of overload. In this second control scheme, the cycle time of the system has to be kept in hand. As the product of queue length and service time is in this case higher than the expected cycle time, it is necessary to construct a fixed value tracking control which can ensure this. This can be done by installing an integral term and by using output feedback. A further benefit of integral control is that as the output disturbance appears on the output, it can also take care of its suppression.

As there is no task in this model for the stabilization of the queue length variable and it can be excluded from the point of view of control, there are only two state variables in the new model: \(x_{\text{proc}}\) and \(x_{\text{serve}}\). The integral term, however, has to be integrated into the system model:

\[
x_i(n + 1) = x_i(n) + y_i(n) = x_i(n) + Cx(n)
\]  

(2.22)

The state matrices take up after the introduction of expanded state variable \(\hat{x} = (x_{\text{proc}} x_{\text{serve}} x_i)^T\) the following form:

\[
\begin{align*}
\hat{x}(n + 1) &= \hat{A}\hat{x}(n) + \hat{B}\hat{u}(n) \\
\hat{y}(n) &= \hat{C}\hat{x}(n), \text{ where}
\end{align*}
\]  

(2.23)

(2.24)

\[
\hat{x} = \begin{pmatrix} x_{\text{proc}} \\ x_{\text{serve}} \\ x_i \end{pmatrix}, \quad \hat{u} = (u_{\text{serve}}), \quad \hat{A} = \begin{pmatrix} 0 & c_{\text{serve}} & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \hat{C} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}
\]

The poles of this new system are also not stable, inherently because of the newly introduced integral term, though as can be seen on (2.25), the controllability nature remained
and so the design of a feedback control is possible.

\[
\text{rank } (\mathcal{C}(\hat{A}, \hat{B})) = \text{rank } \left[ \hat{B} | \hat{A}\hat{B} | \hat{A}^2\hat{B} \right] = \dim(\hat{A}) = 3 \quad (2.25)
\]

The new system’s state feedback vector can be derived again from Ackermann’s formula (2.16), the new feedback vector is:

\[
\hat{K} = \begin{pmatrix}
\frac{1 - \alpha_1 - \alpha_2 + \alpha_1\alpha_3 + \alpha_1\alpha_2 + \alpha_2\alpha_3}{c_{\text{serve}}} \\
\frac{1 - \alpha_1 - \alpha_3 - \alpha_2 - \alpha_1\alpha_3 + \alpha_1\alpha_2 + \alpha_2\alpha_1}{c_{\text{serve}}} \\
\end{pmatrix}
\]

(2.26)

Here, the feedback of \(x_{\text{serve}}\) can again be excluded by the observation of the following equation.

\[
\alpha_2 = 1 - (\alpha_1 + \alpha_3) \quad (2.27)
\]

So, the simplified vector \(K\) is the following:

\[
\hat{K} = \begin{pmatrix}
\frac{-\alpha_1\alpha_3 - \alpha_1^2 - \alpha_3 + \alpha_3^2}{c_{\text{serve}}} & 0 & -2\alpha_1\alpha_3 + \alpha_1^2 - \alpha_3^2 + \alpha_1^2\alpha_3 + \alpha_1\alpha_3^2 \\
0 & 0 & \frac{0.32}{c_{\text{serve}}} & 0 & \frac{0.064}{c_{\text{serve}}} \\
\end{pmatrix}
\]

(2.28)

By choosing an appropriate stable pole triplet as an example:

\[
\alpha_1 = \alpha_3 = 0.8; \alpha_2 = -0.6 \quad (2.29)
\]

Vector \(K\) of the integral control becomes the following:

\[
\hat{K} = \begin{pmatrix}
0.32 \\
0.064 \\
\end{pmatrix}
\]

(2.30)

### 2.4.5.2 Control synthesis

In the two previous sections, two controls of different philosophical background have been elaborated for two different operation conditions. If the system operates properly, stabilization is easy, while in the case of overload a tracking control has been constructed in order to limit cycle time.

The situation becomes even more complicated because an inconstant parameter - which by its nature has an influence on cycle time - can be found in the system and so its effect has to be taken into consideration. Thus in order to achieve the full synthesis of control, the following two basic tasks have to be performed:

- A subsystem is needed which is able to provide an adequate estimation or measuring about service time parameter \(c_{\text{serve}}\), and
- A block which realizes the switch between control architectures.

Fig. 2.5 shows the block schema of the realized system, where three main blocks represent the separated functions:
The plant with the server system handling the queue and generating processor load,

• The load estimator to calculate the unknown portion of the processor load, and

• The control block.

Determining parameter $c_{serve}$ - as it will be shown later - does not demand full precision but it is also not negligible. For this the measurable actual CPU load and the $x_{serve}$ number of the services can be used. Simple regression methods, such as moving window regression would seem to be the simplest solution though it could lead to an inaccurate result. The other possibility is model-based the estimation of CPU load. By using the original system’s state equations (2.42) the - service independent - CPU load becomes the following:

$$d_{proc}(n+1) = c_{cico}(d_{clo}(n) + d_{cli}(n)) + c_{ib}d_{ri}(n) + c_{ekp}x_{cln}(n)$$

(2.31)

The parameters in the equation ($c_{cico}, c_{ib}, c_{ekp}$) can considered to be nearly constant. According to the system’s characteristics, it is true that:

$$x_{proc}(n+1) = d_{proc}(n+1) + c_{serve}x_{serve}(n)$$

(2.32)

So, the subsystem has to be built up so, that it has to be true for parameter $c_{serve}$, which is a variable in it, that:

$$x_{proc}(n+1) - d_{proc}(n+1) - c_{serve}x_{serve}(n) \rightarrow 0$$

(2.33)

For the minimization of difference, it is suitable to use approaching of $c_{serve}$ with slow integrator, and creating the difference function by output feedback. The feedback gain can be chosen for an appropriate setting time and so the structure becomes a low-pass filter for the parameter.

Fig. 2.6 shows the schema of built-in control. In order for us to be able to choose between the two control strategy, the queue length parameter has to be examined. It is important that the switch between the controls should not be bound to a concrete value (barrier of overload) but it is necessary to introduce hysteresis to a certain degree.
2.4.6 Robustness

When dealing with a system with known parameter uncertainty the question of robustness arises. Although other factors that affect the dynamics, be they parameters that considered in control design or in load estimation can be well defined and having low deviation, the dynamic of request serving ($c_{serve}$) can show high differences as mentioned before. This indicates that this parameter has to be investigated in terms of robustness, defining the confidence interval where the closed loop controls remain stable. Though there exist classic robustness tests, the used method calculates the poles of the closed loop controls explicitly on the basis of deviation from $c_{serve}$.

The problem can be originated from the fact that the feedback is calculated from the optimal value of $c_{serve}$ while in the real system this parameter can not be considered constant. Though this value can be estimated with a proper low pass filter, as described in the previous section and could be utilized as a reconfiguring parameter of the feedback this is not necessary in this case. For investigating robustness, an uncertainty parameter ($\Delta M$) is introduced to give the relationship between the real and the optimal value of $c_{serve}$:

$$c_{real} = \frac{c_{serve}}{\Delta M}$$ (2.34)

Hence the poles of the closed loop can be calculated as the roots of the characteristic polynomial of the system:

$$\phi_c(z) = det \left( zI - (A - BK) \right)$$ (2.35)

In case of the first control scheme of the system (2.14) with the defined poles in the example (2.20) and the design feedback (2.21) the characteristic polynomial can be written as:

$$\tilde{\phi}_c(z) = det \left( zI - (\tilde{A} - \tilde{BK}) \right) = det \begin{bmatrix} z & -\frac{c_{serve}}{\Delta M} & 0.0 \\ 0.024 & 0.0 \\ 0 & 1.0 & z - 1.0 \end{bmatrix}$$ (2.36)
Leading to:

\[ \hat{\phi}_c(z) = z^3 - 1.0z^2 + \left(0.256 + \frac{0.024}{\Delta_M}\right)z - \frac{0.024}{\Delta_M} \]  \hspace{1cm} (2.37)

The poles therefore can be easily determined as a function of the deviation \(\Delta_M\). The condition of stability is to keep their values within the unit circle on the complex plane, which is ensured for the values of the example in case \(\Delta_M > 0.02781\) as can be seen on Fig. 2.7.

The same method can be used for the second control structure of system (2.25) with the given poles of the example (2.29) and the calculated feedback (2.30). The characteristic polynomial can be determined as:

\[ \hat{\phi}_c(z) = det \left(zI - \left(\hat{A} - \hat{B}K\right)\right) = det \left[ \begin{array}{ccc} z & -\frac{c_{acc\,mp}}{\Delta_M} & 0 \\ \frac{0.32}{\Delta_M} & z & 0.064 \\ -1 & 0.0 & z - 1 \end{array} \right] = \]  \hspace{1cm} (2.38)

\[ = z^3 - z^2 - \frac{0.32}{\Delta_M}z + \frac{0.384}{\Delta_M} \]  \hspace{1cm} (2.39)

And so the poles can be determined explicitly again. For the previously defined poles the system remains stable when \(\Delta_M > 0.38\). The visualization of results can be examined on Fig. 2.8.

Naturally the confidence intervals are dependent on the previously defined poles of the control schemes though they can be easily calculated and can support the proper determination of the feedback based on the required confidence.
2.4.7 Validation and simulation results

The chapter outlined the design of a switching control structure that performs the regulation of a server architecture represented by a M/G/1 queue model and a service model according to the following aspects. The primary aspect of construction is to ensure the system’s availability by controlling cycle time with service number per cycles, the second aspect is the shortening of the queue, the serving of incoming requests.

The solution presented in the previous section is easily implementable and what is important from practical aspects the realization of the queue control does not require high resources from the processor.

For the examination of control, a theoretical load model has been used that includes two situations. The two main factors having influence on the performance of the system is the service time of each request ($c_{serve}$) and the rate of the incoming tasks. Fig.2.9 shows the simulated values of these parameters.

First, a cold start scenario is simulated, where the clients are connecting after a short server outage. This can happen in real environment and can be caused even by network breakdown or by system update. In this case the clients are connecting rapidly and having buffered requests that are need to be served. This raises the rate of incoming tasks in a cycle until the server catches up with the accumulated jobs. The effect of this phenomenon can be investigated on the beginning of Fig.2.9(a). After the clients can post their requests to the queue the incoming rate reduces to normal level.

The second setting is a major increase in request service time as can be seen in Fig.2.9(b). Between cycles 600 and 1200 the value of $c_{serve}$ dramatically increases that leads to the problem that the cycle time can not be kept even with regular amount of incoming requests. Naturally this scenario can not become permanent though the increased load of the server hardware (let it be eventual or periodically scheduled) could lead to this scenario. A random noise is also added to the alteration of the value to
emulate the minor random changes in the factor.

Figure 2.9: Simulated service time and incoming request for control validation

Naturally under regular conditions the server could maintain the requests normally with the firstly introduced control scheme. Though in both settings the queue begin to grow, leading to the state were the request can not continuously be served in the cycle time (Fig. 2.10(a)) and the hysteresis switch controller turns the system to the second control scheme to keep cycle time (Fig. 2.10(b)).

Figure 2.10: Queue length and control schemes used for the actual state

The estimation of $c_{serve}$ is a key point for the controller to define the adequate control, the output of this subsystem can be seen in Fig. 2.11.

Under the adequate decision about the control scheme to be used the system generates the number of requests to be fulfilled ($x_{serve}$) to satisfy the goal of the current situation. This can be investigated on Fig.2.12(a). Under regular conditions the dynamics of the control is calm, while in the two extreme situations the second control scheme adapts to the changes in the conditions. When there is a sudden change in the situation the control overshoots and a peek value can be observed for a few cycles. This
is inevitable though has a minor effect as it has been stated before. The final control goal, the keeping of the cycle time can be followed in Fig.2.10(b). It can be seen that despite the necessary overshoot at the start of the second setting at cycle 600, the control maintains the cycle time at the desired level.

Figure 2.11: Estimation of $c_{\text{serve}}$

Figure 2.12: The control input ($x_{\text{serve}}$) and the resulting processor load (cycle time)
2.5 New results: Thesis I

I have designed a control method for the communication subsystem of a telemonitoring system based on the needs of rail transportation represented as a queueing model, taking its load and service time into account. The method controls the linear constrained system model switching between different state-feedback control schemes based on the estimated load.

Related publications: [Aradi, 2007], [Aradi and Bécsi, 2008], [Bécsi and Aradi, 2008], [Bécsi and Aradi, 2010], [Bécsi and Aradi, 2009]

- I have defined the roles of the interface subsystem responsible for the communication with the vehicles as a result of the examination of the IT architecture of the railway management systems.
- I have designed a communication model structure able to serve the complex needs of the railway telemonitoring systems, in particular the needs of reliability and modularity.
- I have pointed out the importance of the performance-control of the server application considering the limited computational resources.
- I have elaborated the linear state space representation of the system which, considering random arrival process and stochastic service time, acts as a M/G/1 type queueing model. The discrete linear form of the model is the following:

\[
\begin{align*}
x(n+1) &= Ax(n) + Bu(n) \quad \text{(2.40)} \\
y(n) &= Cx(n) \quad \text{(2.41)}
\end{align*}
\]

\[
x = \begin{pmatrix} x_{\text{proc}} \\ x_{\text{serve}} \\ x_{\text{ql}} \\ x_{\text{cln}} \end{pmatrix}, \quad u = \begin{pmatrix} u_{\text{serve}} \\ d_{ri} \\ d_{clo} \\ d_{cli} \end{pmatrix},
\]

\[
A = \begin{pmatrix} 0 & c_{\text{serve}} & 0 & c_{\text{ekp}} \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & c_{ib} & c_{\text{cko}} & c_{\text{cko}} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}
\]

\[
C = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}
\]

- I have designed two control schemes for the model (Figure 2.5), one for the keeping of the cycle time of the service for overloaded situations, and another for normal operation. I have chosen a model-based estimation of service time based on continuous measurement and filtering. The switching between the different control schemes utilizes a hysteresis rule for the estimated load value.
Chapter 3

Development of Vehicle On-board Communication System for Rough Environment

3.1 Introduction

Utilizing the already-installed train interconnection cables as the physical layer of an extended intra-train communication could be a cost-effective solution. However these interconnection solutions are not optimal for the standardized digital data transfer solutions. This chapter gives a brief summary on the theoretical aspects of data transmission, and experimental test results of the digital data transfer using non-standard physical layers. Through the example of intra-train networks where the need for extending communication possibilities is present with the growing need of on-line telemetry, theoretical solutions are presented, which have been validated with laboratory and field measurements.

On-line communication, telemetry and fleet management gained a significant role in today’s modern rail transport. Such systems utilize a high level of integration of communication, data management and control systems. A significant part of these systems is the intra-train communication network. To extend the information chain on train level, a network is needed to reach all units of a train. Multiple train communication networks exist for general train control purposes, such as remote traction control to handle push-pull train operation, door and light control, or audio channels. However the need for extended services of intra-train communication has arisen. This need serves several purposes:

- Such systems can provide train specific data to the passenger information system improving passenger satisfaction by providing real-time information about the state of the journey, i.e. current delay, estimated arrival, connections etc. This information could be displayed in passenger cars via various kinds of displays, which obtain data from the on-board unit of the driver’s stand. Other information, such
as seating availability could be shown on an external display. In case of freight trains the system could improve the tracking of trade items.

- Another benefit is the collection of train information for the operator. The telemetry or the automatic enrolment of train units can be centralized on train level with a closed communication solution, where the only connection to a central data centre of the operator is managed by the on-board unit (OBU) of the driver’s stand.

- The system can improve the safety of railroad operations by the remote monitoring of various mechanical elements of a railcar. [Edwards et al., 2005]

Naturally there exists a standardized extension of the Train Communication Network [Zeltwanger, 2012] [Zeltwanger, 2010] in newly introduced passenger and traction units providing additional channels and protocols for these purposes, however the spread of these solutions are low since the life-cycle of railway units are quite long compared to road transport, and the operators need to handle the problem on units already purchased and intended to be in service for a long time.

This motivates the operators to implement such communication systems on their current fleet [Russo et al., 1993]. The task can be handled on two different ways from which the first is to install additional wiring and connectors on their units. This solution could result in a well designed multi-purpose and the most important closed and independent communication system; however the installation costs are high. The other way is to use the already installed connection system of the trains. Naturally one has to examine the pin allocation of the connectors of the current system and define to use those wire pairs that does not carry safety-critical information in order to keep the safety level of the train operation.

First, the specialities of the vehicular environment are presented followed by the brief introduction of the transmission line theory and the differential signalling technologies. As a conclusion of these experimental laboratory validation of the non-standard media proves the feasibility of the theorem. Implementing such system comes with additional problems, such as network build-up, network lookup and enumeration of units. For these problems an algorithmic solution is presented. At the end of the chapter the final field verification of the thesis group is given.

### 3.2 Vehicular environment

The topology of the train units and the formation of the already installed interconnections indicate that such a communication system could only use two kinds of network topology, the chain or the bus topology, as shown on Figure 3.1. Since chain topology breaks the continuity of the cable, bus topology seems to be the suitable solution for the proposed network solution.

The problem becomes more complex, since the Hungarian State Railways uses four different systems for train interconnection three old standards left from the Comecon era,
and the UIC 558 [MAV, 1999] standard. All systems have different wiring, shielding, and termination carrying multiple kinds of analogue or digital signals on high or low voltage and current. These parameters influence the feasibility of the communication.

As an example application of UIC 558 type 18 pin interconnection is examined. The function allocation of the UIC 558 wire is shown on Table 3.1 while wire groups and the standardized connector is shown in Fig. 3.2. Though theoretically one could superpose digital signal transfer to any wire of the interconnection, wires 9-18 transfer safety critical functions therefore the utilization of the audio-related wires are desired.

Moreover the network topology of the train is not fixed, since the individual units can build up different formations in length, or in the position of the traction and/or control units. This feature of the modular network effects in several problems. The proper or even adaptive termination a wire pairs, the determination of achievable network speed, and the network build up have to be solved. Another problem is the safe separation of communication devices in case the train uses the interconnection wire for its primary function. The discussion of these is topics - although they are an important aspect of the design of such system - is not intended in my research, but the examination of the
Table 3.1: Function allocation of the UIC 558 cables

<table>
<thead>
<tr>
<th>Wire Pair</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3 (-)</td>
<td>4</td>
</tr>
<tr>
<td>3 (+)</td>
<td>4 (+)</td>
</tr>
<tr>
<td>3 (+)</td>
<td>4 (-)</td>
</tr>
<tr>
<td>5 (+)</td>
<td>6 (-)</td>
</tr>
<tr>
<td>7 (+)</td>
<td>8 (-)</td>
</tr>
<tr>
<td>7 (X)</td>
<td>7 (Y)</td>
</tr>
<tr>
<td>9 (+)</td>
<td>12 (-)</td>
</tr>
<tr>
<td>10 (+)</td>
<td>12 (-)</td>
</tr>
<tr>
<td>11 (+)</td>
<td>12 (-)</td>
</tr>
<tr>
<td>14 (+)</td>
<td>12 (-)</td>
</tr>
<tr>
<td>15 (+)</td>
<td>12 (-)</td>
</tr>
<tr>
<td>16 (+)</td>
<td>12 (-)</td>
</tr>
<tr>
<td>17 (X)</td>
<td>18 (Y)</td>
</tr>
<tr>
<td>S</td>
<td>Shield of wires 17-18</td>
</tr>
<tr>
<td>13</td>
<td>Common shield for all wires</td>
</tr>
</tbody>
</table>

3.3 Transmission line principle

The term 'transmission line' in electromagnetics is commonly reserved for those structures which are capable of guiding Transverse Electromagnetic (TEM) waves. Transmission-lines are a special class of the more general electromagnetic waveguides. TEM waves can only exist in structures which contain two or more separate conductors. Coaxial lines, parallel plates, and two-wire lines are examples of practical transmission-lines [York A., 2013]. Practically a transmission-line is a two-port network (see Figure 3.3) connecting a generator circuit at the sending end to a load at the receiving end.

Figure 3.3: General transmission line [Hui Tat, 2011]
One can analyse the transmission lines using circuit theory concepts breaking the line into small sections so that the circuit element dimensions will be much smaller than the wavelength. To do this, we describe the transmission-line by a series resistance per unit length $R$, series inductance per unit length $L$, shunt conductance per unit length $G$, and shunt capacitance per unit length $C$. A small section of transmission-line with length $dz$ thus has the following equivalent circuit as shown on Figure 3.4. This concept is also known as distributed parameter representation.

![Figure 3.4: Transmission line section [York A., 2013]](image)

The following parameters are related to the physical properties of the material filling between the wires:

$$LC = \mu \varepsilon \quad \frac{G}{C} = \frac{\rho}{\varepsilon} \quad (3.1)$$

where $\mu$, $\varepsilon$, $\rho$ are the permittivity, permeability, conductivity of the insulator of the wires respectively. The characteristic impedance $Z_0$ is the most important parameter of the transmission line. It depends on the distributed parameters (which depend on the material of the conductor and the surrounding of the wires) and angular frequency $\omega$ but not the length of the line.

$$Z_0 = \frac{R + j\omega L}{\gamma} = \frac{\gamma}{G + j\omega C} = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \quad (3.2)$$

For lossless transmission lines $R = G = 0$ and it can be proved that the frequency dependence of the characteristic impedance will cease thus we can approximate $Z_0 \approx \sqrt{L/C}$.

### 3.3.1 Reflections in transmission lines

In subsequent analyses, we will consider only lossless transmission lines. Consider a finite length transmission (see Figure 3.5) terminated with a load impedance $Z_L$ at the end.
At the position of the load \((l = 0)\) the voltage is \(V_L\) and the current is \(I_L\) thus we have:

\[
V_0^+ + V_0^- = V_L \tag{3.3}
\]

\[
\frac{V_0^+}{Z_0} - \frac{V_0^-}{Z_0} = I_L \tag{3.4}
\]

\[
\frac{V_L}{I_L} = Z_L \tag{3.5}
\]

where \(V_0^+, V_0^-, I_0^+, I_0^-\) are the wave amplitudes in the forward and backward directions at \(z=0\) and \(\gamma\) is the complex propagation constant given by \(\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)}\). By solving the above equations we obtain:

\[
\Gamma_L = \frac{V_L}{V_0^+} = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{3.6}
\]

And finally by summarizing the special cases we end up in the following results:

- Short circuit: \(Z_L = 0 \rightarrow \Gamma_L = -1\)
- Open circuit: \(Z_L = \infty \rightarrow \Gamma_L = +1\)
- Matching termination: \(Z_L = Z_0 \rightarrow \Gamma_L = 0\)

From the practical point of view we can state that in case the data rate is low or the cables are short termination may be unnecessary. As data rates increase, termination becomes important. Since any device on the bus can transmit, it is probable that a node within the middle of the bus will transmit requiring that termination be applied to both ends of the bus segment. The termination impedance have to match the characteristic impedance of the wire. Using high frequency communication the imaginary part of the impedance is negligible thus a normal resistor (termination resistor) can be used.

### 3.4 Differential signalling technologies

During the review of the possible solutions the constraints of the vehicular environment have to be considered which fundamentally affect the technological possibilities.
• Given the train interconnections as transmission line, which consists of shielded and twisted wire fours.

• The maximum bus length is 500 m.

• Bus topology shall be used.

• Rough environment (extreme temperature, humidity and mechanical stress).

• Standard communication technology shall be used.

Based on the above the solution could be one of the differential signalling technologies. Two of the collision domain segment network buses can satisfy all of these requirements: EIA-485 and Controller Area Network.

3.4.1 TIA/EIA RS-485-A

One of the more popular technologies for interconnecting devices on a network is TIA/EIA-485-A, known throughout industry as RS-485, see [Electronic Industries Association, 1998]. According to the standard it specifies the characteristics of the generators and receivers used in a digital multipoint system. It does not specify other characteristics such as signal quality, timing, protocol, pin assignments, power supply voltages or operating temperature range.

An EIA-485 bus usually consists of two or more communication controllers each powered by a separate power source. At a minimum, a single shielded or unshielded twisted-pair cable interconnects the various controllers in a daisy-chain fashion. In some instances a short stub is allowed; however, higher speed networks usually do not allow stubs. Star topology is definitely not recommended. Termination is usually applied to the ends of the network. The standard basically specifies the parameters (unit load, output drive, common mode voltage etc.) of the drivers, receivers and transceivers.

![Figure 3.6: EIA-485 network structure](image)
attached to the network. Basically a driver must be able to source at least 1.5 volts differentially into 60 ohms (two 120 ohm terminators in parallel along with 32 unit loads) under a common mode voltage range of -7 to +12 Vdc. Data rates are not specified and there are a wide range of devices that conform to the standard but are intended either for high speed (up to 50 Mbps) or low speed (skew rate limited). In terms of the Open Systems Interconnection Reference Model (OSI), EIA-485 only defines the lowest layer, the physical layer. It is used by several higher layer protocols such as Profibus, ARCNET and other token-based protocols. There are several key topics that must be considered when deploying EIA-485 networks such as termination, fail-safe bias, connectors, grounding, cabling and repeaters. From our point of view the most important parameters are the termination and the cabling. Terminating a data cable with a value equal to its characteristic impedance reduces reflections that could cause data errors. The most popular approach is DC termination although this approach results in higher power dissipation. Resistive terminators typically have values of 120 to 130 ohms although twisted-pair cable impedances can be as low as 100 ohms. An 100 ohm termination resistor is too low for the EIA-485 drivers. A value closely matching the cable impedance must be applied at some convenient location closest to the ends of the cable segment as possible.

3.4.2 Controller Area Network

The Controller Area Network (CAN) [Robert Bosch GmbH, 1991] is a serial communications protocol which efficiently supports distributed real-time control with a very high level of security. Its domain of application ranges from high speed networks to low cost multiplex wiring. In automotive electronics, engine control units, sensors, anti-skid-systems, etc. are connected using CAN with bitrates up to 1 Mbit/s.

CAN is a multi-master bus with an open, linear structure with one logic bus line and equal nodes. The number of nodes is not limited by the protocol. Physically the bus line (Figure 3.7) is a twisted pair cable terminated by termination A and B. Locating the termination within a CAN node should be avoided because the bus lines lose termination if this CAN node is disconnected from the bus line. The bus is in the recessive state if the bus drivers of all CAN nodes are switched off. In this case the mean bus voltage is generated by the termination and by the high internal resistance of each CAN node’s receiving circuitry. A dominant bit is sent to the bus if the bus drivers of at least one unit are switched on. This induces a current flow through the termination resistors and, consequently, a differential voltage between the two wires of the bus. The dominant and recessive states are detected by transforming the differential voltages of the bus into the corresponding recessive and dominant voltage levels at the comparator input of the receiving circuitry [ISO, 2003].

The CAN standard [ISO, 2003] gives specification which shall be fulfilled by the cables chosen for the CAN bus. The aim of these specifications is to standardize the electrical characteristics and not to specify mechanical and material parameters of the cable. Furthermore the termination resistor used in termination A and termination B shall comply with the limits specified in the standard also.

37
Beside the physical layer the CAN standard specifies the ISO/OSI data link layer also. CAN uses a very efficient media access method based on arbitration principle called "Carrier Sense Multiple Access with Arbitration on Message Priority" [ISO, 2003]. Summarizing the properties of the CAN network we can declare the following according to the CAN specification [Robert Bosch GmbH, 1991]:

- Prioritization of messages
- Guarantee of latency times
- Configuration flexibility
- Multicast reception with time synchronization
- System wide data consistency
- Multi-master
- Error detection and signalling
- Automatic retransmission of corrupted messages as soon as the bus is idle again
- Distinction between temporary errors and permanent failures of nodes and autonomous switching off of defect nodes

These properties result in significant advantages against the EIA-485 standard.

### 3.5 Laboratory Experiments

As mentioned in the introduction the first step of using such non-standard communication should be the experimental verification of feasibility under laboratory conditions. It should be assumed that the ends of the network can not be terminated because in

![Figure 3.7: CAN bus structure](ISO, 2003)
this case the operator would have to switch the termination resistors at both ends of the network during every train composition. Therefore every active communication node must have an individual ‘termination’. Though these resistors can not have a standard resistance because more than two nodes may be connected to the bus and in this case the equivalent resistance of the parallel resistors may be too small causing the overload of the transceivers. In these laboratory tests the effect of the non-standard termination resistor with different cable lengths and data rates are examined on a CAN bus.

Table 3.2: Laboratory test results

<table>
<thead>
<tr>
<th>Length</th>
<th>Termination (Ω)</th>
<th>Baud rate</th>
<th>Error rate</th>
<th>Cable type</th>
</tr>
</thead>
<tbody>
<tr>
<td>10m</td>
<td>-/-</td>
<td>250K</td>
<td>0%</td>
<td>UTP Cat. 5</td>
</tr>
<tr>
<td>10m</td>
<td>-/-</td>
<td>500K</td>
<td>100%</td>
<td>UTP Cat. 5</td>
</tr>
<tr>
<td>10m</td>
<td>120/-</td>
<td>1M</td>
<td>0%</td>
<td>UTP Cat. 5</td>
</tr>
<tr>
<td>20m</td>
<td>-/-</td>
<td>125K</td>
<td>0%</td>
<td>UTP Cat. 5</td>
</tr>
<tr>
<td>20m</td>
<td>-/-</td>
<td>250K</td>
<td>100%</td>
<td>UTP Cat. 5</td>
</tr>
<tr>
<td>20m</td>
<td>120/-</td>
<td>1M</td>
<td>0%</td>
<td>UTP Cat. 5</td>
</tr>
<tr>
<td>40m</td>
<td>-/-</td>
<td>100K</td>
<td>0%</td>
<td>UTP Cat. 5</td>
</tr>
<tr>
<td>40m</td>
<td>-/-</td>
<td>125K</td>
<td>100%</td>
<td>UTP Cat. 5</td>
</tr>
<tr>
<td>40m</td>
<td>120/-</td>
<td>500K</td>
<td>0%</td>
<td>UTP Cat. 5</td>
</tr>
<tr>
<td>40m</td>
<td>120/-</td>
<td>1M</td>
<td>100%</td>
<td>UTP Cat. 5</td>
</tr>
<tr>
<td>40m</td>
<td>120/120</td>
<td>1M</td>
<td>15.17%</td>
<td>UTP Cat. 5</td>
</tr>
<tr>
<td>80m</td>
<td>-/-</td>
<td>100K</td>
<td>0%</td>
<td>FTP Cat. 6</td>
</tr>
<tr>
<td>80m</td>
<td>-/-</td>
<td>125K</td>
<td>100%</td>
<td>FTP Cat. 6</td>
</tr>
<tr>
<td>80m</td>
<td>120/-</td>
<td>250K</td>
<td>0%</td>
<td>FTP Cat. 6</td>
</tr>
<tr>
<td>80m</td>
<td>120/-</td>
<td>500K</td>
<td>0%+15 errorframe</td>
<td>FTP Cat. 6</td>
</tr>
<tr>
<td>80m</td>
<td>120/120</td>
<td>500K</td>
<td>0%</td>
<td>FTP Cat. 6</td>
</tr>
<tr>
<td>80m</td>
<td>120/120</td>
<td>1M</td>
<td>100%</td>
<td>FTP Cat. 6</td>
</tr>
<tr>
<td>80m</td>
<td>1500/-</td>
<td>100K</td>
<td>0%</td>
<td>FTP Cat. 6</td>
</tr>
<tr>
<td>80m</td>
<td>1500/-</td>
<td>125K</td>
<td>100%</td>
<td>FTP Cat. 6</td>
</tr>
<tr>
<td>80m</td>
<td>1500/1500</td>
<td>125K</td>
<td>0%</td>
<td>FTP Cat. 6</td>
</tr>
<tr>
<td>80m</td>
<td>1500/1500</td>
<td>250K</td>
<td>100%</td>
<td>FTP Cat. 6</td>
</tr>
</tbody>
</table>

The effect of the order of magnitude of higher termination resistors were tested with different cable lengths and baud rates. The test setup was composed of standard UTP cable \((Z_0 = 120\Omega)\) with switchable termination resistors at both ends and USB-CAN interfaces. During the measurements 10 different messages with extended ID and 8 data-bytes were sent with a suitable frequency causing approx. 100% bus load. In every test cases 10 000 messages were sent from the one end and logged on the other end of the network. In Table 3.2 the termination type and value is signed:

- -/-: open circuit,
- R/-: RΩ termination at one end,
• R/R: \( R \Omega \) termination at both ends.

First, totally unshielded twisted pair (U/UTP) cables were used with different bus lengths, then shielded twisted pair (F/UTP) were used with fixed length. The measurements found that at lower baud rates the bus can work without termination, but using only one termination resistor can cause significant improvement in the communication. Of course the measurements verified that the highest baud rates can be achieved with standard termination at both ends.

Fig.3.8. shows the results of the test. The measurement clearly proved that the CAN network can operate with non-standard termination, though on lower bandwidth.

![Figure 3.8: Laboratory test, achievable CAN speed with different termination and length](image)

Figure 3.8: Laboratory test, achievable CAN speed with different termination and length

The most important result related to the non-standard termination resistors. In this case 1500 ohm resistors (weak termination) were used which results in a properly high equivalent resistance with more than two nodes. These tests found that the maximal baud rate decreased to quarter of the standard values, but it is twice as the open circuit values. These tests were performed with 1200 ohm resistances with the same results. It is also remarkable that in a borderline scenario the CAN protocol detected the errors (15 error frames) and resent the messages. The aim of these tests was the verification of the behavior of such a CAN bus. The laboratory experiments found that the network with non-standard termination can operate with twice higher data rate as the open circuit.

### 3.6 Network look-up algorithm

As mentioned before, knowing the order of the units in the train is essential since the proper positioning of the termination in the CAN network highly increases the achievable
bandwidth. On the other hand the determination of the order of the devices in the network also serves operational purposes by enabling the proper automatic enumeration of the train’s units. For this task a proper network lookup algorithm is needed.

Because of the given conditions the train has bus network topology for the CAN communication as it can be seen in Fig.3.9. It can be seen that all on-board units (OBU) of the train are connected to a two-wire bus provided by the interconnection cables. The problem is that the topology is not fixed, since the composition of the train can be freely manipulated by the operator.

Figure 3.9: Proposed network topology

To enable the determination of the order of the units the interconnection cable has a circuit breaker beside all OBUs. Since the circuit breaker also prevents the original function of the wire pair from operation, it has to be designed under high safety standards (e.g. using fail-safe relays) and the ability to close whenever the original functionality is needed. These switches are normally in closed state, and only open during the execution of the network lookup algorithm.

The base concept behind the algorithm is the following: At start all CAN devices utilize weak termination (for example 1500Ω) enabling low bandwidth (1 kbit/s) on the network even with relatively long bus. With this bus speed the number, position and orientation of each unit can be determined. With this knowledge the network can be properly terminated at the ends of the train and the all OBUs can switch to higher bandwidth.

Since the media access of the CAN network is based on arbitration there is no master role on the network. Thus to control the network lookup algorithm a 'master' OBU has to be chosen for this task. This way the communication on the network has the following stages (see Fig.3.10.):

- Normal communication
- Network Lookup
- Disconnected (Media used by original function)

Since the wire pair chosen for the communication is rarely used, the normal state is considered as all switches are closed, and the OBUs are communicating on the network. This state can be interrupted by two ways: one is when the original function needs the media, and the other is when a new OBU appears on the network.

When the original function needs the media, the OBUs leave their switches closed, but disconnect their transceivers from the network.
When a new unit ID appears on the network the network lookup algorithm starts. First, all OBUs send their unique identifiers on the network. The one with the lowest value is chosen as the master for communication.

During the network lookup all units including the master continuously sends their IDs on the network, and their order is determined by their visibility depending on the state of the circuit breakers.

The master systematically sends a message to each OBU (including itself) to temporarily open its circuit breaker. This message is available for anyone on the network, so all OBUs know which unit opens the media separating the network into two halves. For this short time all OBUs receive only the IDs of those units which are on the same side of the network. This way all OBUs will have a set of visible IDs for each opened breaker.

Fig. 3.11. shows a simple example for this systematic enumeration. The figure shows the algorithm from the point of view of the left side unit with two additional units having different orientation. It can be seen that with any combination of the orientation of the two units their visibility is different according to the states of the switches and so both their order and orientation can be determined. After acquiring all set of visible OBUs for all opened switches, every OBU can determine the order and orientation with two easy steps:

1. First it determines the orientation of each OBU by examining the set generated when the given units circuit breaker was opened. If the unit’s ID is in this set it means that its tranceiver is closer than its circuit breaker.

2. Second the unit excludes all visible unit IDs from their sets, and orders the set with any sorting algorithm that gives the order of the units. Naturally with OBUs
in the middle of the train can have to handle the units differently on their two sides.

Table 3.3 shows an example with 9 unit train from the view of the left side end of the network. For the simplicity of the example the OBUs order and their IDs are identical and the "Spin" column shows the orientation of the unit. Spin value 0 means that the circuit breaker is on the left of the CAN transceiver and 1 means it’s on the right. It is shown that only the IDs of the units with the spin of 1 (3,4,8,9) are appearing in their corresponding sets. By excluding these identifiers the cardinality of set determines the distance from the left end.

Let us define the set of all visible identifiers as $A$ and the set of the IDs from the point of view of the unit with the ID $i$ when the circuit breaker of the $j$th unit is opened as $A^i_j$. If we state that the circuit breaker of the $i$th unit is on the left of its transceiver than the units on its right side are:

$$U^{right}_i = A^i_i$$  \hspace{1cm} (3.7)

And the units in its left:

$$U^{left}_i = A \setminus A^i_i$$  \hspace{1cm} (3.8)
Table 3.3: Example with 9 unit train from the view of the left side end of the network

<table>
<thead>
<tr>
<th>Spin</th>
<th>ID Opens</th>
<th>Visible IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1,2,3*</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>1,2,3,4*</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>1,2,3,4</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>1,2,3,4,5</td>
</tr>
<tr>
<td>0</td>
<td>7</td>
<td>1,2,3,4,5,6</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>1,2,3,4,5,6,7,8*</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>1,2,3,4,5,6,7,8,9*</td>
</tr>
<tr>
<td>None</td>
<td></td>
<td>1,2,3,4,5,6,7,8,9</td>
</tr>
</tbody>
</table>

The determination of the orientation of any unit depends on its position according to unit $i$ as can be seen on Table 3.4:

Table 3.4: Determination of orientation of each unit

<table>
<thead>
<tr>
<th>$j$ ∈ $U^\text{left}_i$</th>
<th>$j$ ∈ $U^\text{right}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$j$ ∈ $A^j_i$</td>
<td>Right (1)</td>
</tr>
<tr>
<td>$j$ /∈ $A^j_i$</td>
<td>Left (0)</td>
</tr>
</tbody>
</table>

The order can be determined from the distance of each unit, which can be formally written as:

$$D^j_i = \begin{cases} 
|A^j_i \setminus U^\text{left}_i \setminus \{j\}| & \text{if } j \in U^\text{right}_i \\
|A^j_i \setminus U^\text{right}_i \setminus \{j\}| & \text{if } j \in U^\text{left}_i 
\end{cases}$$

An other issue is the switching between different data transfer rates on the CAN network. Generally in case when one transceiver with different bandwidth setting connects to a CAN network it automatically detects the messages as erroneous, disrupts the data transfer and starts to send error frames to the network. This undesired phenomenon can be avoided by using 'listening' or 'listen-only' mode of the CAN devices [Koppe, 2003]. After start-up the OBUs can initiate their CAN controller to listen-only mode and listens alternately in the two possible baud rates and detects the speed of the network. If no communication is presents it starts with low bandwidth.

Special cases exist when network connection and disconnection events occur:
1. If any disturbance occurs on the network resulting in erroneous network operation, the devices can detect it and fall back to listening/low bandwidth mode.

2. When one or more unit disconnects, their periodic identifier stops appearing on the network. At this case, the device with the currently highest ID priority sends a signal to perform the network lookup.

3. The same algorithm occurs when new identifiers appear on the network.

4. When two working networks are connected the termination resistance halves because of their parallel configuration. This is the operational boundary of the CAN network. Depending on the current conditions this can be resulting in either erroneous or correctly working network. If the communication on the network remains, this case leads to case 2. When this connection ruins the communication it leads to case 1.

5. When the train units are disconnected the termination could be ineligible leading to case 1.

6. When a device restarts, it stops sending its identifier for a short period of time. If the identifier reappears in a previously defined time there is no need for network lookup.

### 3.7 Experimental results

To examine the behaviour of the concept in real conditions a field test was organized with the support of the Hungarian State Railways. The train was formed by 10 Bhv type coaches and a Bdt type driving trailer, with all units having the same length of 23.740m. The driving trailer was placed in the middle of the train and the units were connected by 18 pole UIC558 type interconnection cables.

Ten units were involved in the test all having switchable 1500 Ohm resistors, and two 120 Ohm switchable resistor was installed in the 1st and 10th unit. The wire pair used for the communication was:

- 5. Switch on loudspeaker amplifiers (+) for CANH
- 7. Priority of announcement command (+) for CANL

Two CAN analysers were installed in the 1st and 10th car to generate and measure network load. The distance between the two can devices was approximately 230 meters. CAN Frames with extended IDs and 8 byte length data field were sent on the network, with different data values, for example (0x55) for alternating bits and (0x00) for long unchanged state, where bit-stuffing occurs. The network load was set to be at least 60%.

Table 3.5. summarizes the measurement results. Same measurements were made between the 1st and 2nd and the 1st and 5th units. It can be said as a general conclusion that CAN technology is robust enough for communication purposes via train
interconnection, even with high length and non CAN-standard cabling and termination. Based on the measurements it can be assumed that CAN communication with 20-50kbit/s bandwidth can be safely applied on existing train interconnection solutions of the railways.

Table 3.5: Field measurement, CAN network with different termination, using UIC 558 interconnect cable

<table>
<thead>
<tr>
<th>CAN Termination</th>
<th>Bus speed</th>
<th>Network Operable</th>
<th>Average Bus Load</th>
<th>Percentage of erroneous frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>1500Ω every unit</td>
<td>100</td>
<td>+</td>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td>1500Ω every unit</td>
<td>125</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>120Ω at two ends</td>
<td>125</td>
<td>+</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>120Ω at two ends</td>
<td>250</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>120/1500Ω at two ends</td>
<td>125</td>
<td>+</td>
<td>68</td>
<td>0.019</td>
</tr>
<tr>
<td>1500Ω at two ends</td>
<td>20</td>
<td>+</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>1500Ω at two ends</td>
<td>50</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Though the problems arising from this method are diverse, since safety constraints of the existing communication media, proper design of the non-standard physical layer of the CAN communication, the switching between network states and the handling of the network control and enrolment algorithm all need proper approach the benefits that can be obtained from this solution is also high.

Beside of the current railway-specific application, the measurements of the non-standard CAN network hold individually separable results, since the utilization of CAN communication in any other field with electrical constraints can be widespread.

The laboratory and field measurements proved the feasibility of the solution from the view of achievable stable data transfer. It can be said that this solution is feasible with quite long train units without using any repeaters or signal amplifiers.

The network lookup algorithm presented in this chapter is fast enough for the needs of this application, since the slower part of it, when the circuit breakers are operated has linear $O(n)$, whilst the sorting algorithm afterwards has quadratic $O(n^2)$ (worst-case) complexity.
3.8 New results: Thesis II

I have proposed a design of an intra-train network for vehicles equipped with a telemonitoring system, which considers the environmental disturbances and the variable network topology, ensuring a reliable communication channel between the on-board units.

Related publications: [Aradi et al., 2013a],[Aradi et al., 2015b]

• I have examined the different digital communication technologies capable of transferring data on long-distance wired media and chosen the one, which seems to be the most appropriate for the inspected environment.

• I have defined the effects of the non-standard media and termination resistance values on the performance of the Controller Area Network (CAN) - the selected technology. For this task I have conducted laboratory measurements, which confirmed that the technology is reliable under the modified conditions, but on lower data transfer rate.

• Reproducing the measurements described above on trains I have validated that the prototype system based on the proposed method works reliably under the noisy environment conditions of rail transportation.

• For the automated determination of train composition and also for the proper handling of network terminations I have developed an algorithm for the case of variable topology CAN network.
Chapter 4

Design of Predictive Optimization Method for Energy-Efficient Operation of Trains

4.1 Introduction

In the previous chapters the data acquisition issues and extension possibilities of the online telemonitoring system were detailed and several solutions were proposed considering the railway requirements. This chapter focuses on the utilization of the collected data in the field of vehicle control.

A new method to estimate the train propulsion resistance parameters using the online telemonitoring data is introduced. A new approach to the design of speed profile for trains concerning timetable and energy-efficient operation is proposed. The developed method takes into consideration the inclinations of the railway tracks, the speed limits, the non-linear tractive effort curve and resistances to achieve the energy-efficient speed. It is demonstrated that by using a predictive optimization approach both punctuality and reduction of energy consumption can be achieved. The proposed method has been tested by simulation based on a real case study. The presented algorithm could be used in drivers’ training or as a core algorithm for automatic train operation.

The introduction of the proposed method is organized as follows. The first part of the chapter starts with Section 4.2, which details my motivation and the applied research method, Section 4.3 presents the energy saving potential in railway operation confirmed with real train data, Section 4.4 describes and formalizes the energy consumption of a train and Section 4.5 proposes a novel method to estimate the parameters of the train propulsion resistance. The next topic deals with the optimization and control starting with the generation of an optimal reference run in Section 4.6. In the following, Section 4.7 the predictive optimization method and algorithm is presented, Section 4.8 shows the simulation results with real data compared to experimental results of the Swiss railway and finally Section 4.9 presents sensitivity and robustness of the proposed method.
4.2 Motivation

The development of an energy-efficient operation strategy has been in the focus of research and development centers, suppliers and industrial/rail users. It is motivated by the competition between rail and road transportation, in which effectiveness, operation costs and reliability are important factors as mentioned formerly. The problem of minimising the traction energy consumption can be split into three main tasks: choosing the proper vehicle type, development of energy-optimal timetable and optimal control of the train, see [Zobory, 2010].

The design of energy-efficient traction leads to various numerical optimization methods. In the first papers a large number of simplifications were assumed in order to develop methods which were fast enough even with the available computational power, see e.g., [Ishikawa, 1968; Milroy, 1981; Strobel and Horn, 1973]. Some examples of the simplifications are the linear expression of resistances, constant constraints, the assumption that the external forces do not depend on the train position, and that the inclinations and the tracks are constant between two stations.

Several experimental tests have shown that the inclinations have significant effects on energy consumption, see e.g., [Meyer et al., 2007]. In order to handle changing slopes Maksimov proposed a method first in 1971, but only for underground trains with short station distance, see [Maksimov, 1971]. Golovitcher et al. developed an energy optimal control for rail vehicles moving along a known route. Using the maximum principle, they found a set of optimal controls, the control switching graphs and complementary conditions of optimality [Liu and Golovitcher, 2003]. The traction force was formalized as a continuous variable, which was suitable for modelling an electric locomotive, but not a diesel one.

The discrete traction force model for optimal control of diesel freight trains was taken into consideration in the research of Howlett et al., see [Howlett, 1990; Howlett et al., 1994; Howlett, 1996]. They developed various optimal driving strategies for on-board control. Using the local energy minimization principle they calculated the critical switching points for a global optimal strategy, see [Howlett et al., 2009].

In another important approach only the coasting point was used for actuation and the coast control of train movement was proposed, see [Wong and Ho, 2003, 2004a,b]. They used meta-heuristic (genetic) algorithms for computing the proper coasting point offset. Using these solutions the consideration of the changing speed restrictions was strongly limited. The railway network with its interactions was considered in an energy saving train operation as a disturbance condition, see [Fu et al., 2009; Yang et al., 2012]. They assumed the predetermined routing and traversing order plan and used the coasting control method to seek optimal control strategies.

Several driver advisory systems exist on the market both for trains (such as Knorr-Bremse Leader) or for suburban railway or trams (such as in Dresden), for more details see [Aradi et al., 2014b]. But all of these systems requires an individual on-board unit and sensor system. In the last decade more and more traction companies have introduced on-line telemonitoring systems as mentioned in Chapter 2, that can provide the necessary
information and computation performance for implementing driver advisory systems or automatic train operation (ATO) functions. On the one hand it is motivated by the deployment of new traffic safety and management systems such as the European Rail Traffic Management System/European Train Control System (ERTMS/ETCS), on the other hand the increasingly widespread operation control and information systems can also provide the necessary infrastructure for a driver advisory system.

The objective of my research is to develop an energy-efficient optimization method using the data of the on-line telemonitoring system which can be used in driver assistance systems. Another purpose is to achieve lower computational requirements, since the method should be capable of real-time operation on low-performance on-board computers. The ultimate task, however, is to create the basis of the autonomous control system. A new method is developed for the design of a predictive optimal control. The input data are the track profile, the required trip time, traction and braking force characteristics, speed limits, slopes, which are available in an on-board database (updated regularly via the on-line communication system). The quality criterion is the energy consumption and the state variables are the train position and speed. The optimization itself is based on the sequential quadratic programming method, in which the calculation time is also taken into consideration.

4.3 Energy Saving Possibilities

In this section the result of a statistical analysis is introduced in which the rate of the possible energy savings are calculated using the on-line telemonitoring system data. During the analysis one month real train data was processed from the MÁV Electronic Logbook System, narrowing the data to the electric locomotives. Having regard to the exact energy consumption comparing, nearly identical engine types were chosen, i.e. the 431, 432 and 432 locomotive series of MÁV.

4.3.1 General Analysis

First a general statistical analysis is introduced. The purpose of this analysis is to show the deviation of the energy consumption of trains running under the same conditions using a sufficiently large sample. In addition the analysis can show the effect of the human factor on the energy consumption.

In order to compare the individual train runs properly the train load should be taken into consideration, i.e. to evaluate the real energy consumption of the runs the effect of mass parameter should be taken into account. In accordance the load-consumption value pairs were generated. Figure 4.1 shows the correlation between the energy consumption and train load using real train data. It is important to calculate the theoretical consumption as a function of the train load.

The regression possibilities were examined and the linear regression was chosen regarding the preliminary analysis of the samples, assuming that average energy consump-
tion is the function of the train load:

\[ y = \alpha x + \beta, \text{where} \]

- \( x \) is the train load;
- \( y \) is the corresponding energy consumption; and
- \( \alpha, \beta \) parameters.

To determine \( \alpha \) and \( \beta \) the well-known equations of the linear regression were used, where the two parameters are the solutions of the following system of equations,

\[
\begin{align*}
\alpha \sum x_i^2 + \beta \sum x_i &= \sum x_i y_i \quad (4.2) \\
\alpha \sum x_i + n\beta &= \sum y_i, \quad (4.3)
\end{align*}
\]

where \( n \) is the number of the sample points.

From this point the theoretical energy consumption is considered as the above mentioned linear regression, i.e. it determines the average consumption of a real train run. The relative deviation of the real and theoretical consumption can be calculated with the following equation:

\[
E_{rel}(m) = \frac{E_{re} - E_{th}(m)}{E_{th}(m)}, \text{where} \]

- \( E_{rel} \) is the relative deviation of a given run;
- \( E_{re} \) is the measured energy consumption; and
- \( E_{th}(m) \) is the calculated theoretical consumption as a function of train mass.
The first question is whether there are significant differences between the consumptions of each run with the same conditions, i.e., is there any energy saving possibilities in the system. For a high level examination of the question, the statistics of the relative deviation of the energy consumption of approximately 270000 individual runs is generated in such a way that every train run was split into smaller runs between two standstill positions. The results can be seen on the histogram presented in Fig. 4.2(a). The horizontal axis shows the relative differences in percentage grouped in classes of 5 percent intervals, while the vertical axis shows the relative frequency of the runs for each class. The cumulative probability of the absolute deviation is presented in Fig. 4.2(b). It can be seen that the 90% of the consumption values are located inside the 20% deviation. Thus the real consumption values are distributed in a 40% area. It proves the existence of significant energy saving possibilities. It should be noted that these statistics do not take the traffic circumstances, the weather and track conditions into consideration, though it will be proved that significant savings can still be achieved just on focusing on the driver behaviour.

![Histogram](a) Histogram ![Cumulative consumption](b) Cumulative consumption

Figure 4.2: The energy consumption differences of the inspected journeys

### 4.3.2 Analysis of Driving Style

As mentioned above, significant deviations exist between the train runs. However, it is an important question whether only the traffic situation or also the human factor plays a role in the energy consumption. From the point of view of the driver advisory system, it is a very important aspect. For this purpose, the same dataset were analysed as above, but with considering the differences between the engine-drivers. As the deviation of the real and theoretical energy consumption for each run were generated previously, thus this data can be grouped by other properties, such as the identifier of the engine-driver. The dataset was narrowed before the analysis resulting in that only the data of those engine-drivers was taken into consideration who has got at least 100 records in the dataset, furthermore the runs with more than 30% relative deviations were filtered also as their cause of extreme consumption were considered as independent from driver behaviour.
The steps of the processing are the following:

1. Remove the runs with more than 30% relative difference in consumption \((0.7 < E_{rel} \text{ or } 1.3 > E_{rel})\).

2. Group the reduced dataset by the engine-drivers and filter to those having more than 100 records.

3. Generate the average energy consumption for each engine-driver.

4. The drivers are classified into 1% interval groups by their relative consumption deviation.

At the end of this process 1005 engine-drivers have matched the restrictions. The diversity of the drivers can be seen in Fig. 4.3(a), and their cumulative distribution function is shown in Fig. 4.3(b). It can be seen in the figures that the distribution of consumption deviation is smaller than in the above mentioned general statistics, but even significant.

![Histogram](image1)

![Cumulative consumption](image2)

(a) Histogram  
(b) Cumulative consumption

Figure 4.3: The energy consumption differences of the individual engine drivers

These analyses have proven that significant deviation exists between the consumption of the analysed train runs. For these results the runs has to be analysed on a common base, namely the energy consumption per unit load. The analyses have shown significant deviation in the consumption statistics, which is caused on one hand by the human factor. This insists that the driving styles of the drivers in the Hungarian rail transport is diverse and it has significant effect on the energy consumption. On the other hand the driving-style-independent deviation is larger, since other effects should exist coming from timetable editing or traffic incidences.

4.3.3 Section Analysis

After the general and driving-style analysis the research has continued with the analysis of the sections. For this purpose the No. 80 Miskolc-Nyékládháza section was chosen
since it has a relatively busy schedule table and also data amount and quality is suitable. The stops at the inspected subsection can be seen in Table 4.1 and the simplified map is shown in Figure 4.4. The analysis has involved 725 individual train runs on the mentioned section which is a sufficiently large sample. Another advantage of this section that the correlation between the energy consumption and the trains’ gross mass is high as can be seen in Figure 4.1.

Table 4.1: Stations at the studied section of Line 80.

<table>
<thead>
<tr>
<th>Station</th>
<th>Section (100m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miskolc-Tiszai</td>
<td>1817</td>
</tr>
<tr>
<td>Miskolc Rendező</td>
<td>1802</td>
</tr>
<tr>
<td>Kistokaj</td>
<td>1734</td>
</tr>
<tr>
<td>Nyékládháza</td>
<td>1689</td>
</tr>
</tbody>
</table>

Figure 4.4: Map of Line 80 between Miskolc and Nyékládháza

During the analysis the effect of several parameters on the consumption were examined. Figure 4.5 shows influence of the total journey time, the net journey time (total time decreased with the stop time) and the number of stops during the journey. The diagrams and the regression lines shows that these three indicators (which were previously assumed as correct) do not correlate with energy consumption. The problem is that the same journey time could be achieved with smooth speed profile and with aggressive acceleration-deceleration also.

Because the energy lost by braking and reused by acceleration plays significant role in the consumption growth, the number of decelerations and the lost velocity (practically kinetic energy) correlate with the energy consumption growth. It is shown in Figure 4.6. According to the mentioned conditions the analysed section was performed with 0-4 decelerations by all drivers, where the definition of the deceleration is a velocity changing
process where there is at least 10 km/h decrease in velocity followed by an acceleration stage of at least 10 km/h speed gain. The figures show the clear connection between the number of decelerations and the energy consumption growth. In this case the key question is what causes the decelerations. The reasons could either be traffic conditions and human factors. Each of them can be eliminated or at least decreased by an adequate driver advisory system.

The analysis has confirmed the well-known fact, that number and scale of the decelerations highly affect the energy consumption of trains. But another important aspect is the differences between the train runs with nearly identical speed profiles caused by driving-style. So it can be stated that two basic factors influence the traction energy consumption. Firstly the traffic circumstances, where the negative effects are caused by the delays and secondly the human factor, namely the driving-style. If the main goal is the improvement of the system, the traffic circumstances should be considered as an inevitable negative effect and it should be eliminated locally. Naturally the task is a multi-criteria optimization, since besides the minimal energy consumption the keeping of the journey time is the aim also. However such a system may not disturb the basic task of the engine-driver. These goals can be reached with the development of the
following three parts of the system.

- Education
- Motivation
- Advisory system

The first is the education of the engine-drivers. In this task the engine-driver should learn the effect of the deceleration and acceleration and energy demand of the chosen speed. Another important thing is the practising the different traffic use cases and sections in simulators. Probably experienced engine-drivers possess this knowledge, but they are not motivated to use it. So the second pillar is the motivation of the engine-drivers, comparing their performance with a statically or dynamically (in a traffic-dependent way) calculated energy consumption norm. The motivated engine-drivers could be supported with a driver advisory system. In the further sections newly developed driver advisory algorithms are shown based on the above mentioned on-line telemonitoring system data.

4.4 Energy Consumption of Rail Vehicles

4.4.1 Problem description

As mentioned above the task is the optimization of the train’s movement between two stations with minimum energy consumed considering the restrictions. The result of the optimization have to be an optimal speed profile (speed over position function). This problem could be detailed as follows.

The train should move between two stations (stops) with minimum traction energy $E$ consumed considering the trip time $T$ given for a section with length $S$. In addition the model must satisfy some restrictive conditions:

- The traction force $F_{trac}$ must not exceed either the $F_{trac}^{max}(v)$ maximum speed-dependent traction force capability $F_{eng}^{max}(v)$ of the modelled locomotive engine or the maximum adhesion force $F_{adh}^{max}(v)$:

$$F_{trac}(v) \leq F_{trac}^{max}(v) = \min \left( F_{eng}^{max}(v), F_{adh}^{max}(v) \right) \quad (4.5)$$

- Keeping the speed limits during the journey must be considered as hard constraint since the train’s speed must never exceed the limits given in the actual rail section:

$$0 \leq v(s) \leq v_{lim}(s), \quad \forall s \in (0, S) \quad (4.6)$$

- Keeping time at the end of the journey can be formalized as follows:

$$t(S) = T, \text{ where } t(0) = 0 \quad (4.7)$$
The brake system of the modern railway vehicles consists of two main subsystems: an electric (regenerative) brake and a frictional (pneumatic) brake. Usually they are used together as a blended electro-pneumatic brake system [Miyatake and Ko, 2010]. It means that the frictional brake is applied when the regenerative braking force is not sufficient for the specific braking force. The braking force actuated by a frictional brake system can be formulated as an \( F_b = F_b(v, u_b) \) function of two variables, where the independent variables are speed \( v \) and the non-positive (by definition) \( u_b \leq 0 \) control variable [Zobory, 2012]. If the assumption that the frictional coefficient is the exponential function of the speed is excepted the quasi-static braking force can be approximated with the following formula:

\[
F_f(v, u_b) = u_b \frac{u_b}{|u_b_{\text{max}}|} \left( F_{b0} + |F_{b0} - F_{b1}| \cdot e^{-\lambda v} \right) \leq 0, \tag{4.8}
\]

where \( F_{b0} \) is the right-hand-side limit of the braking force function belonging to the \( |u_b_{\text{max}}| \) maximum brake control and \( F_{b1} \) is the limit of the same function in case of \( v \to \infty \). The intensity of the change of exponential functions is given by the \( \lambda \geq 0 \) parameter, which depends on the type and mechanical parameters of the brake system (e.g. tread brake or disc brake) [Zobory, 2012].

The braking force generated by the electric brake system can be described with similar characteristics just as in the case of tractive force and it is also the function of speed and the brake control variable. For acceleration and deceleration, additional constraints may be introduced with regard to passenger comfort. For the scope of this chapter the assumption of optimal adhesion conditions is taken into consideration meaning that the maximum traction force \( F_{\text{trac}}^{\text{max}}(v) \) can be utilized on the entire track.

### 4.4.2 Mathematical formulation

The following differential equations describe the trains longitudinal motion:

\[
\dot{s} = v \tag{4.9}
\]

\[
\dot{v} = \frac{1}{m} (F_{\text{trac}}(v) - F_{\text{brake}}(v) - F_{\text{res}}(v) - F_{\text{slope}}(s)) \tag{4.10}
\]

where \( s \) and \( v \) are the position and speed of the train and the considered forces affecting its movement are the following: \( F_{\text{trac}}(v) \) is the traction force, \( F_{\text{brake}}(v) \) is the braking force, \( F_{\text{res}}(v) \) is the propulsion resistance and \( F_{\text{slope}}(s) \) is the force caused by inclinations. (see Fig.4.7.)

Many research projects deal with the longitudinal resistance forces that affect the train [Iwnicki, 2006; Rochard and F., 2000; Profillidis, 2000] and it can be said that all handle the propulsion resistance \( F_{\text{res}} \) consistently as a second order polynomial of the train speed:

\[
F_{\text{res}}(v) = \alpha + \beta v + \gamma v^2 \tag{4.11}
\]
where $\alpha, \beta, \gamma$ are constants. Though there exist some generalized methods for these coefficients calculating with aerodynamic coefficients, axle count and axle load and some other parameters, the general practice is to determine the coefficients by curve fitting on experimental measurements for each type of traction unit. The forms of propulsion resistance equations used and the empirical factors selected vary between railway systems reflecting the use of equations that more closely match the different types of rolling stock and running speeds. Propulsion resistance comes from different sources, according to Hay [Hay, 1961]:

1. Internal resistance of the locomotive- from cylinder and bearing friction, power used by auxiliaries such as lighting, heating, air compressor, etc.

2. Resistance varying directly as the axle loading, called journal friction.

3. Resistances that vary directly with the speed and is called flange resistance, such as flange friction, oscillation, swaying etc.

4. Resistance varying approximately with the square of the speed, called air resistance;

5. Track modulus resistance;

6. Wind resistance

Though there exist many parameters and equations for the propulsion resistance of longitudinal train dynamics (see Fig.4.8 and Table 4.2), the identification of actual parameters for a chosen locomotive should be advantageous.

$K_a$ is an adjustment factor depending on rollingstock type; $k_a d$ is an air drag constant depending on car type; $m_a$ is mass supported per axle in tonnes; $n$ is the number of axles; $V$ is the velocity in kilometres per hour; and $\delta V$ is the head wind speed, usually taken as 15 km/h.

The external force $F_{slope}$ caused by the inclination of the track is the following:

$$F_{slope}(s) = \delta(s)F_G$$

(4.12)
Table 4.2: Empirical Formulas for Propulsion Resistance, Freight Rollingstock [Iwnicki, 2006]

<table>
<thead>
<tr>
<th>Description</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified Davis equation (U.S.A.)</td>
<td>$K_a[2.943 + 89.2/m_a + 0.0306V + 1.741k_{ad}V^2/(m_a n)]$</td>
</tr>
<tr>
<td>French Locomotives</td>
<td>$0.65 m_a n + 13n + 0.01 m_a n + 0.03 V^2$</td>
</tr>
<tr>
<td>French Standard UIC vehicles</td>
<td>$9.81(1.25 + V^2/6300)$</td>
</tr>
<tr>
<td>French Express Freight</td>
<td>$9.81(1.5 + V^2/(2000...2400))$</td>
</tr>
<tr>
<td>French 10 tonne/axle</td>
<td>$9.81(1.5 + V^2/1600)$</td>
</tr>
<tr>
<td>French 18 tonne/axle</td>
<td>$9.81(1.2V^2/4000)$</td>
</tr>
<tr>
<td>German Strahl formula</td>
<td>$25 + k(V + \Delta V)/10k$</td>
</tr>
<tr>
<td>Broad gauge (i.e., 1.676 m)</td>
<td>$9.81[0.87 + 0.0103V + 0.000056V^2]$</td>
</tr>
<tr>
<td>Broad gauge (i.e., 1.0 m)</td>
<td>$9.81[2.6 + 0.0003V^2]$</td>
</tr>
</tbody>
</table>

Figure 4.8: Comparison of propulsion resistance equations [Iwnicki, 2006]
where $\delta(s)$ is the slope in the specific position, $F_G$ is the force of gravity.

Thus, the optimization objective is to find an optimal speed profile $v(s)$ for $(0 \leq s \leq S)$, to minimize the cost function of energy consumption:

$$E_0^S = \int_0^S F_{trac}(v(s))ds \to \min,$$  \hfill (4.13)

while satisfying (4.5), (4.6) and (4.7).

4.5 Parameter Estimation

In order to determine running resistance, an appropriate experimental method must be chosen. This algorithm should eliminate as many error sources as possible while keeping the cost of the tests at minimum. There are a number of methods whose purpose is to determine rolling resistance and aerodynamic drag. Generally they form three main distinct groups [Lukaszewicz, 2001, 2007]:

- Tractive effort methods.
- Dynamometer or drawbar methods.
- Coasting methods.

In the first method the measurements are taken with the train under traction. The propulsion resistance is calculated by measuring the power produced by the traction system. It is important to have high accuracy in the inclination data of the track. These measurements should preferably utilize constant speed, since accelerations highly modify power consumption and thus has to be corrected. Energy efficiency of the engine is also has to be taken into consideration. Another option is to measure the torque transferred to the wheel rims by strain gauges, which eliminates the need for the efficiency estimation of the engine.

To measure the resistance of only the rolling stock, or the internal losses of a tractive unit the second method utilizes a dynamo-meter, placed between a winch and a cable connected to the train or vehicle. By pulling the train smoothly on a straight track with constant gradient the examined parameters could be obtained. In this case, acceleration and retardation also must be measured and compensated for. This method is generally resource consuming, and the precision and usability of the achieved information is questionable.

With the utilization of the third method the errors and disturbances in the measurement of the tractive energy can be eliminated. The test needs a track section with known altitude gradients. First the train accelerates to a previously defined speed, and by reaching the start of the section it begins to coast, meaning no tractive or braking effort is applied afterwards. From the measurement of the speed loss one could calculate
the loss of kinetic energy which is the sum of the rolling resistance and the potential energy loss or gain resulting from the inclinations of the track. Naturally in this case the effect of the wind should be considered somehow.

4.5.1 Proposed method

As mentioned in the previous section, the determination of the propulsion resistance equation’s parameters is needed for the calculation of energy consumption of the train. The proposed method used for the parameter estimation could be classified into the first method group with some special additions.

As an illustration of the algorithm, freight rolling stock with the Hungarian State Railways’ 0431 series locomotive (formerly named as V43) was used. Table 4.3 shows the most important parameters of the locomotive.

<table>
<thead>
<tr>
<th>Power type</th>
<th>Electric locomotive (25kV 50Hz AC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Builder</td>
<td>Ganz</td>
</tr>
<tr>
<td>Build date</td>
<td>1963-1982</td>
</tr>
<tr>
<td>Total produced</td>
<td>379</td>
</tr>
<tr>
<td>UIC classification</td>
<td>B’B’</td>
</tr>
<tr>
<td>Gauge</td>
<td>1.435mm (Standard)</td>
</tr>
<tr>
<td>Length</td>
<td>15700 mm</td>
</tr>
<tr>
<td>Locomotive weight</td>
<td>80 t</td>
</tr>
<tr>
<td>Traction motors</td>
<td>2</td>
</tr>
<tr>
<td>Top speed</td>
<td>120 km/h</td>
</tr>
<tr>
<td>Power output</td>
<td>2200 kW</td>
</tr>
</tbody>
</table>

The base idea behind the process is to determine the parameters without conducting any specific measurements so only the data provided by the telemonitoring management system was used. Five individual freight trains were chosen which were running on the same section, between Budapest-Kelenföld and Tatabánya (for overview of the track see Figure 4.9).

To perform parameter estimation one need to know some fundamental informations about the circumstances and conditions about the investigated runs:

- Weight and type of the traction unit
- Number, weight and preferably the axle count of the rolling stock
- Speed and energy consumption measurements as a function of travel time and distance
- Inclinations of the track
The locomotive on-board computers measure and record position, speed, primary voltage, current and phase with one second sampling time. Thus real power consumption can be calculated with the well known formula.

Position is given in WGS84 (World Geodetic System) coordinates by the GPS receiver, and speed is given from two sources, from the GPS and from the locomotive’s own speedometer. According to these data and knowing some characteristic points of the track, the error of the longitudinal position calculations could be kept under 0.1%. This low error level in this data is essential, since it has to be synchronized with the altitude diagram of the track which is given as a function of longitudinal distance on the line.

The altitude diagram of the 60 km long section used for the process, the speed trajectory of the trains, the energy consumption and the calculated tractive forces are shown on Figure 4.10, while the basic parameters of can be examined in Table 4.4.

All tractive units and rolling stock are of the same type, though there are uncertainties and unknown disturbances in these measurements. One is that the gross mass of the trains is given as a result of cargo mass calculation and can contain 5-10% error. The strength and direction of the wind is unknown, and the mechanical condition of the wagons, rims and the locomotive is also unknown, therefore the parameter estimation should consider these phenomena as an error factor in mass and the other parameters as some kind of cumulated efficiency parameter.

According to the researches and formulas presented in the previous section the propulsion resistance is divided into two parts, one is for the locomotive ($F_{\text{loc res}}$) and one is for the rolling stock ($F_{\text{stock res}}$). The parameters are specifically defined for an actual wagon and locomotive type, thus the formula of the resistance forces can be formed as:
Figure 4.10: Track elevation, speed trajectory, tractive force and energy consumption of sample runs used for parameter estimation

<table>
<thead>
<tr>
<th>Running Time</th>
<th>Gross mass (t)</th>
<th>Wagon Count</th>
<th>Wagon mass (t)</th>
<th>Energy consumption (kWh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>48’30’</td>
<td>557</td>
<td>24</td>
<td>20</td>
<td>893</td>
</tr>
<tr>
<td>47’10’</td>
<td>510</td>
<td>19</td>
<td>22</td>
<td>694</td>
</tr>
<tr>
<td>45’31’</td>
<td>644</td>
<td>11</td>
<td>51</td>
<td>835</td>
</tr>
<tr>
<td>44’30’</td>
<td>1243</td>
<td>19</td>
<td>61</td>
<td>1208</td>
</tr>
<tr>
<td>45’40’</td>
<td>882</td>
<td>17</td>
<td>47</td>
<td>1104</td>
</tr>
</tbody>
</table>
\[ F_{\text{loc}}(v) = m_{\text{loc}} \ast (\alpha_l + \beta_l v) + \gamma_l v^2 \] (4.14)

\[ F_{\text{stock}}(v) = m_{\text{stock}} \ast (\alpha_s + \beta_s v) + n_{\text{stock}} \ast \gamma_s v^2, \] (4.15)

where \( v \) is the speed in m/s, \( m_{\text{loc}} \) is the weight of the locomotive, \([\alpha_l, \beta_l, \gamma_l]\) are the parameters of the locomotive resistance, \( m_{\text{stock}} \) is the gross mass of the rolling stock, \( n_{\text{stock}} \) is the number of wagons and \([\alpha_s, \beta_s, \gamma_s]\) are the coefficients of the rolling stock’s resistance equation. The overall propulsion resistance of the train is the sum of these two forces:

\[ F_{\text{res}}(v) = F_{\text{loc}}(v) + F_{\text{stock}}(v) \] (4.16)

To determine these parameters based on the measurements having one second sample time, some simplifying assumptions were made:

- Track slope, and acceleration is considered as constant for the one second sample. Since the dynamics of the railways is slow, this assumption is a minor simplification.

- The consumed energy is used in the same time interval. This simplification could cause much more problem since the inner delays in the power and transmission are greater than the sample time.

Parameter estimation was made by modeling the energy balance of the trains in each sample time which is consist of kinetic- and potential energy and the work generated by the resistance force. Without disturbances and uncertainties the energy consumption of the \( j^{th} \) train in the \( i^{th} \) sample is the following:

\[ \Delta \bar{E}_{i,j} = \frac{1}{2} m_j (v_{i,j}^2 - v_{(i-1),j}^2) + \Delta h_{i,j} m_j g + F_{\text{res},i,j} \Delta s_{i,j}, \] (4.17)

where \( \Delta \bar{E}_{i,j} \) is the energy balance of the sample, \( \Delta h_{i,j} \) is the altitude change and \( \Delta s_{i,j} \) is the length of the inspected section. As mentioned above the uncertainty of the mass must be considered with a multiplicative parameter for each train’s rolling stock therefore the real considered mass of the units after the locomotive is:

\[ m_{j}^{\text{stock}} = \xi_j (m_j - m_j^{\text{loc}}), \] (4.18)

where \( \xi_j \) is the mass uncertainty of the whole train and having the value range of [0.9, 1.1]. Mechanical efficiency is considered, when the energy balance is positive, thus the locomotive needs additional tractive effort:
\[ \Delta \hat{E}_{i,j} = \begin{cases} \psi_j \Delta \hat{E}_{i,j} & \text{if } \Delta \hat{E}_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (4.19) \]

where \( \psi_j \) is the constant aggregated efficiency of the train during the whole run with the value range of \([0.0, 1.0]\) and so \( \hat{E}_{i,j} \) is the estimated energy consumption. The vector of unknown parameters are formed as:

\[ x = [\alpha_l, \beta_l, \gamma_l, \alpha_s, \beta_s, \gamma_s, \xi_j, \psi_j]: j \in [1..k], \quad (4.20) \]

where \( k \) is the number of individual trains.

By using equations (4.14)-(4.19) one can calculate the consumed energy of the trains, when the parameters \( x \) are given. To find these parameters, it is necessary to define a fitness function to evaluate the goodness of them. The simplest approach is to determine the sum of square differences of calculated and measured consumption in all sample sections:

\[ g(x) = \sum_{j=1}^{k} \sum_{i=1}^{l_j} \left( \Delta E_{i,j}^{\text{meas}} - \Delta \hat{E}_{i,j} \right)^2, \quad (4.21) \]

where \( l_j \) is the number of samples of the \( j^{th} \) train, and \( E_{i,j}^{\text{meas}} \) is the measured energy consumption. Though this kind of fitness function would enlarge the errors from the neglecting of actuation delay. An other approach should be the comparison of the cumulated energy consumption:

\[ f(x) = \sum_{j=1}^{k} \sum_{i=1}^{l_j} \left( \varepsilon_{i,j}^{\text{meas}} - \hat{\varepsilon}_{i,j} \right)^2, \quad (4.22) \]

where \( \varepsilon_{i,j} \) is the measured and \( \hat{\varepsilon}_{i,j} \) is the calculated cumulated energy consumption of the \( j^{th} \) train till the \( i^{th} \) sample:

\[ \varepsilon_{i,j} = \sum_{h=1}^{i} \Delta E_{i,j} \quad (4.23) \]

According to the equations above, The optimum search problem of the parameter estimation of the propulsion resistance of the train with the given measurements can be formalized as:

minimize \[ f(x) \]

with respect to \[ x = [\alpha_l, \beta_l, \gamma_l, \alpha_s, \beta_s, \gamma_s, \xi_j, \psi_j]: j = 1, \ldots, k \]

subject to \[ 0.9 \leq \xi_j \leq 1.1 \]

\[ 0.0 \leq \psi_j \leq 1.0 \]

\[ x_i \geq 0.0; i = 1, \ldots, 6 \quad (4.24) \]
The optimization problem has only linear equality and inequality constraints, though the fitness function given is continuous nonlinear. By choosing any appropriate minimum search algorithm, such as sequential quadratic programming, trust region reflective or interior point methods, parameters can be determined. On the example set, the actual coefficients of the resistance functions is shown on equations (4.25), (4.26).

\[ F^{loc}_{res}(v) = m_{loc} \left( 0.0121 + 0.0000409v^2 \right) \]  
\[ F^{stock}_{res}(v) = 0.0187m_{stock} + n_{stock} \times 1.1400v^2 \] 

(4.25)  
(4.26)

It can be seen that the optimum search eliminated the linear part of equations (4.14) and (4.15) which is in compliance with most of the previously published formulas.

![Figure 4.11: Measured and calculated (with the found parameters) tractive forces of Train No.1. and 2.](image)

Figure 4.11 shows the comparison of the measured and calculated tractive forces for trains 1 and 2. It should be noted that the measured tractive force is also calculated, it is derived from the measured energy consumption. Unknown disturbances and the effect of the delay can be examined on these sample diagrams.

Figure 4.12 presents the speed profile and the comparison of cumulated energy consumption through the example of train 4 and 5.

### 4.5.2 Error Analysis

The quality of the output of the algorithm may depend on the quality of the data provided by the telematics system. Speed and position measurements, energy consumption or the available track information all may contain errors as it has already been mentioned above. The real question is how sensitive the method is to the errors provided by the onboard measurement system.
Since the algorithm uses numerical optimization, an analytic evaluation of error sensitivity is impossible. Even so it is important to know the correctness of the results given by the method. For this task, the measurement data of the case study presented above was altered to simulate different measurement error scenarios. To demonstrate this, three different scenarios were outlined:

1. Error in the provided track data
2. Error in the measurement of energy consumption
3. Measurement delays

In the first scenario the effect of the inaccuracy in the altitude diagram was examined. For this task one hundred different altitude diagrams were created with the following rules: The long term altitude error – taking the 110 m altitude as base – should not exceed 4%, while the local gradient error should not exceed 10%. The lower and upper bounds of the generated diagrams can be examined in Figure 4.13. The figure also presents the acquired resistance diagrams.

To simulate the inaccuracy in the measurement of the energy consumption the original measurements applied in the case study were altered with a similar rule set to the one described above. The long term error in the energy consumption was kept under 4% and the local gradient error was kept under 10%.

Then with the one hundred generated energy consumption samples the determination of the parameters was also carried out. Sample upper and lower bounds of the altered energy consumption diagram can be examined in Figure 4.14. The different resistance diagrams generated as the result of the altered consumption measurements can also be examined in the figure.
Figure 4.13: Bounds of the altered altitude diagram, and the identified speed resistance diagrams generated with the parameters of Train No.1

Figure 4.14: Bounds of the altered energy consumption diagram of Train No.1, and the identified speed resistance diagrams generated with the parameters of Train No.1
For the emulation of the measurement delays or the asynchronous recordings of the energy consumption and speed/position values, these two sources were shifted in time from -3 seconds to +3 seconds with 0.5 second steps. The resulting resistance diagrams are shown in Figure 4.15.

Figure 4.15: Resistance diagrams of Train No. 1 with asynchronous data recordings

The summary of the results of the error analysis can be seen in Table 4.5. The generated running resistance diagrams were compared to the result of the case study at different speeds taking the train parameters of Train No. 1 as a basis. The three error scenarios were designed to be at least two times worse than the expected error rate. Based on the results of the error analysis the following conclusions can be drawn:

Regarding sensitivity the algorithm showed the worst relative performance at low train speed and the best error suppression was found between 60-80 km/h values. This phenomenon can be explained with the nature of the applied data set. Since more than 80% of the data had a speed value between 60 and 90 km/h the optimization algorithm achieved the minimum of the f(x) fitness function by optimizing the curve to this speed interval. The increasing relative error at higher speed (120 km/h) can be explained the same way.

Table 4.5: Results of the error analysis. The maximum of the absolute and relative error of the generated running resistance diagrams at different speeds (Train No. 1)

<table>
<thead>
<tr>
<th>Speed (km/h)</th>
<th>Error Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Altitude</td>
</tr>
<tr>
<td>0</td>
<td>1.22kN (11.2%)</td>
</tr>
<tr>
<td>60</td>
<td>0.60kN (3.1%)</td>
</tr>
<tr>
<td>120</td>
<td>2.99kN (7.7%)</td>
</tr>
</tbody>
</table>
4.6 Development of the Optimization Algorithm

This section gives two possible solutions for generating reference speed-trajectories for further optimization. The first is a simple model that neglects track inclinations and only considers speed limits, while the second takes track gradients into account. Both algorithms have the purpose of providing a speed profile which is feasible, energy efficient yet keeps the journey time.

4.6.1 Simple reference

This section gives a simple example of generating reference trajectory that only considers speed limits. The simplicity of the algorithm and so the numerical speed is a key point of feasibility. Obviously, when solving a non-linear, multi-criteria optimization one should apply some simplification. The high level approach of the algorithm takes the following steps:

1. Divide the given section into n number of subsections, where the speed limit of the subsections are constant (see Figure 4.16). Naturally the minimal division equals to the distinct speed limits of the section.

![Figure 4.16: Example division of a given section](image)

2. Define a typical speed for all sections \( v_i, i = 1..n \) which will be the holding speed of the section. Beside this define the starting and arrival speed of the train \( (v_0, v_{n+1}) \). Naturally the train stops at the start and at the at of the journey \( (v_0 = v_{n+1} = 0) \). Determine the energy consumption \( (e_i) \) and trip time \( (t_i) \) for each section with a given trajectory pattern considering only three stages: accelerating, holding and decelerating. Thus the algorithm can calculate the overall journey time and energy.
consumption for any $v_i$ vector as:

$$T(v_i) = \sum_{1}^{n} t_i(v_i)$$  \hspace{1cm} (4.27)$$

$$E(v_i) = \sum_{1}^{n} c_i(v_i)$$  \hspace{1cm} (4.28)$$

3. Define an optimum search for the vector $v_i$ to keep the journey time $J = T_i(v_i)$ and minimize $E(v_i)$.

Figure 4.16 shows a division of an example section into five subsections. Each subsection has a distinct speed limit (marked with red line) and having a holding speed ($v_i$).

Table 4.6: The possible speed profile patterns used for the method

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Case</th>
<th>Shape</th>
<th>$v_{start}$</th>
<th>$v_{end}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$v_{i-1} \leq v_i \leq v_{i+1}$</td>
<td><img src="image1" alt="Image" /></td>
<td>$v_{i-1}$</td>
<td>$v_i$</td>
</tr>
<tr>
<td>II</td>
<td>$v_{i-1} \leq v_i \geq v_{i+1}$</td>
<td><img src="image2" alt="Image" /></td>
<td>$v_{i-1}$</td>
<td>$v_{i+1}$</td>
</tr>
<tr>
<td>III</td>
<td>$v_{i-1} \geq v_i \geq v_{i+1}$</td>
<td><img src="image3" alt="Image" /></td>
<td>$v_i$</td>
<td>$v_{i+1}$</td>
</tr>
<tr>
<td>IV</td>
<td>$v_{i-1} \geq v_i \leq v_{i+1}$</td>
<td><img src="image4" alt="Image" /></td>
<td>$v_i$</td>
<td>$v_i$</td>
</tr>
</tbody>
</table>

In case the only available behavior is accelerating, holding and braking, the speed profile of a given subsection can be form four distinct pattern to achieve minimal trip

71
Table 4.7: Kinematic equations of the different phases

<table>
<thead>
<tr>
<th>Phase</th>
<th>Distance Equation</th>
<th>Time Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerating (I,II)</td>
<td>$s_{t}^{acc} = \frac{v_{i}^{2} - v_{i-1}^{2}}{2a_{acc}^{v_{i}}} $</td>
<td>$t_{t}^{acc} = \frac{v_{i} - v_{i-1}}{a_{acc}^{v_{i}}} $</td>
</tr>
<tr>
<td>Braking (II,III)</td>
<td>$s_{t}^{dec} = \frac{v_{i}^{2} - v_{i+1}^{2}}{2a_{dec}^{v_{i}}} $</td>
<td>$t_{t}^{dec} = \frac{v_{i} - v_{i+1}}{a_{dec}^{v_{i}}} $</td>
</tr>
<tr>
<td>Speed holding</td>
<td>$s_{t}^{hold} = s_{t} - s_{t}^{acc} - s_{t}^{dec} $</td>
<td>$t_{t}^{hold} = \frac{s_{t}^{hold}}{v_{i}} $</td>
</tr>
<tr>
<td>Whole subsection</td>
<td>$s_{i}$</td>
<td>$t_{i} = t_{t}^{acc} + t_{t}^{dec} + t_{t}^{hold} $</td>
</tr>
</tbody>
</table>

time. Though this is a conservative approach with high level of simplification, the time keeping property of the algorithm ensures energy gain.

Assuming that the $v_{i}$ holding speed of any subsection is also the maximum speed at the given section since it is preceded with acceleration and followed by braking, the speed pattern of the $i^{th}$ section is depending on the typical speeds of the previous and the following section’s holding speed, as summarized in Table 4.6.

4.6.1.1 Energy consumption with constant acceleration

The energy consumption and journey time of a subsection can be calculated in many different ways. Since the algorithm presented in this section is a preliminary reference generating method, one can make assumptions that the acceleration and deceleration rates are considered constant, even though these are the functions of the traction (braking) force and the train mass. Naturally some numerical differences should appear this way, though the pre-calculating of approximate acceleration rates for any $(v_{i}, v_{i+1})$ speed differences could reduce the error and keeps numerical performance. Table 4.7 summarizes the simple kinematic equations for this case by using the section notations of Table 4.6. Thus, the journey times of the subsections could be easily calculated.

To calculate the energy need of a phase, two factors have to be considered. One is the kinetic energy change, which only applies to the acceleration phases:

$$E_{t}^{mov} = \frac{1}{2} m(v_{i}^{2} - v_{i-1}^{2})$$

(4.29)

The other part of the energy consumption is the work spent on the propulsion resistance $F_{res}$, having the form of:

$$F_{res} = \alpha + \beta v + \gamma v^{2}$$

(4.30)

This work can be calculated as the integral of the resistance force over the distance:
\[ E = \int_{s=0}^{s_1} F_{\text{res}}(v(s)) ds \]  
(4.31)

For speed holding phases this provides the well known simple form of the work:

\[ E_{i\text{hold}}^{\text{hold}} \left|_{a=0} \right. = (\alpha + \beta v_i + \gamma v_i^2) s_i^{\text{hold}} \]  
(4.32)

For accelerating phase the speed function over the distance can be obtained as:

\[ v(s) = \sqrt{2as + v_0^2} \]  
(4.33)

By combining equations (4.30) and (4.33) the tractive resistance can be obtained as a function of distance with known \( v_0 \) initial speed, \( a \) acceleration and the constant parameters of the resistance equation.

\[ F_{\text{res}}(s) = \alpha + \beta \sqrt{2as + v_0^2} + 2\gamma as + \gamma v_0^2 \]  
(4.34)

When resistance force (4.30) is considered as \( \beta = 0 \), this function has the following form:

\[ F_{\text{res}}(s)\left|_{\beta=0} \right. = \alpha + 2\gamma as + \gamma v_0^2 \]  
(4.35)

Integrating (4.34) leads to the energy consumed by the propulsion resistance with the given conditions:

\[ E = \int_{s=0}^{s_1} F_{\text{res}}(s) ds = \frac{1}{3a} \left( -\beta v_0^3 + 3a s_1 a + \beta \left( 2as_1 + v_0^2 \right)^{3/2} + 3\gamma s_1^2 a^2 + 3\gamma s_1 av_0^2 \right) \]  
(4.36)

And also, when \( \beta = 0 \):

\[ E|_{\beta=0} = s_1 \left( \alpha + \gamma s_1 a + \gamma v_0^2 \right) \]  
(4.37)

4.6.1.2 Optimum search

The third step of the method is to find the optimal vector of holding speeds \( (v_i) \) for each subsection to minimize energy consumption and keep the journey time. the optimum search algorithm should handle two objectives this way, the cost function can be formed as the weighted average of the to goals:

\[ J(v_i) = W_e E(v_i) + W_t |T(v_i) - \mathcal{T}|, \]  
(4.38)
The only restriction for the generated $v_i$ vector is that the upper limit of the speed is defined $v_{reg,i}$ for each subsection. The optimization problem can be formalized as:

$$\begin{align*}
\text{minimize} & \quad J(v_i) \\
\text{with respect to} & \quad v_i; \quad i = 1, \ldots, n \\
\text{subject to} & \quad 0.0 \leq v_i \leq v_{reg,i},
\end{align*}$$

(4.39)

where $W_e$ and $W_t$ are weighting parameters and $T$ is the desired journey time.

The values of $E$ and $T$ is given and so the evaluation of $J(v_i)$ is numerically fast, it can not be explicitly calculated and so the minimum search is not trivial for $J(v_i)$. On the other hand the cost function is bounded and continuous therefore numerical optimization methods can easily handle the problem.

The effect of the weighting parameters on the optimization will be discussed later in Section 4.9.2

4.6.1.3 Example

The method presented in the previous section provides a theoretically optimal speed profile for the subsection patterns defined by Table 4.6. To present the behavior of the algorithm, an example is given. The track conditions are presented in Figure 4.17. The section’s length is 31100 meters and the speed limits are shown on the figure.

![Figure 4.17: Example section. Speed limit and speed profile with the lowest journey time](image)

The sample simulations were made with considering a 0431 type locomotive of the Hungarian State Railways, and with a gross train mass of 640 tonnes. The minimum time journey with the given condition takes 18 minutes and 28 seconds.

For the example simulations, Sequential Quadratic Programming (SQP) was chosen as optimum search method. Figures 4.18 and 4.19 present the speed profiles generated
and their corresponding energy consumption diagrams. The test runs where made from the journey time of 18'30 to 21’30 with half minute step. By inspecting the results, it can be seen that the algorithm provides speed profiles analogously to communicating vessels and capillarity in fluid mechanics. As a results, only those speed limit raises inflict acceleration, where the corresponding time gain compensates the additional energy consumption. This phenomenon is best apparent at the speed limit gain at 17000 meters, where only the minimal time journey accelerates to the maximum speed. This simple algorithm is able to provide a conservative yet efficient advice to the train driver, though it is important to mention that the results of the algorithm are only optimal at sections, where track inclination are negligible ie its absolute value is less then 3%.

Figure 4.18: Generated speed profiles with multiple journey times

Since the evaluation of the algorithm is fast, it is able to recalculate the speed profile during the journey, if the driver does not keep its recommendation. Though the major role of the algorithm presented is to become an input to the predictive optimization presented in the latter sections.

4.6.2 Reference considering track inclinations

The method introduced in the previous section searches the answer for the energy-optimal time keeping journey. The algorithm does not utilizes coasting, or any other phase than acceleration, holding and braking. Besides this, the algorithm does not considers the inclinations of the track.

4.6.2.1 Background

As far as the gradient of the track slope does not exceeds the rate, when speed holding results in braking there is not too much energy loss in the method, the suboptimal results
Figure 4.19: Energy consumption of the generated speed profiles with multiple journey times

of the algorithm only arises from the neglecting of coasting behavior, because though the acceleration-holding-braking phases result in less journey time, this does not compensate the energy consumption raise.

Figure 4.20 shows a simple scenario to understand the effect of track inclinations. In this theoretic scenario the track has non-constant gradient \( h \) and the train has to choose its speed profile to cover the distance \( S \) int time \( t(S) - t(0) \) while achieving the speed \( v(S) \) at the end with the initial speed of \( v(0) \). Let us assume, that the journey time of simple speed holding is the time goal of this trip.

The Figure 4.20 shows two possible speed profiles:

- Version "A" shows simple speed holding, and
- Version "B" is a profile optimized for the inclinations.

Both trajectories cover the distance in the same time interval, and the final speed \( v(S) \) is also the same, so the final kinetic energy stored in the systems are the same. Though on the diagram of tractive forces it can be seen that speed holding needs more power to keep velocity in the uphill section while it has to brake on the sweep downhill phase, which is a direct energy loss if there is no regenerative braking. Version "B" on the other hand only utilizes small amount of tractive power to reduce the deceleration in the uphill and then coasts to the end of the section gaining speed from the slopes.

4.6.2.2 Algorithm design

Howlett shows [Howlett et al., 2009] [Howlett and Pudney, 1995] that in constant conditions of track gradient and speed limit, acceleration-coast or acceleration-hold-coast strategies are the best to energy efficiently minimize time of a train’s journey. In
Figure 4.20: Energy consumption of the generated speed profiles with multiple journey times. (Track inclinations, Speed and Forces)

In some cases, where the sweep uphill results in acceleration even in the coasting phase, acceleration-coast-hold strategy should be used. At the end of the section where the train must reduce its speed under a certain speed limit braking may occur.

Figure 4.21 shows some of these possibilities with the same starting speed, and speed limit on the section and at the section. The difference between the cases is the track gradient. It can be seen, that there are two cases:

- acceleration-hold-coast-brake and
- acceleration-coast-hold-brake

Naturally in some cases parts of these sequences can be omitted.

The idea behind the generation of a reference speed profile that considers inclinations is that if one can calculate these profiles for given conditions explicitly, the optimization scheme introduced in section 4.6.1 can be extended.

In this case the full journey has to be divided into more smaller subsections since not only the speed limit, but the track gradient has to be constant (with some error naturally) on the whole subsection.
#### 4.6.2.3 Kinematic calculations

The previous algorithm gained speed from the simple calculation of the subsections’ kinematics, the goal is to find the explicit calculation for a given subsection with given starting speed, limits, gradient and of course journey time. The following calculations try to find effects of speed dependent propulsion forces at constant tractive force rates, and gives the method of constructing the strategies introduced above.

As shown in (4.10) the following differential equation describe the trains longitudinal motion:

\[
\frac{dv(t)}{dt} = \frac{1}{m} \left( F_{\text{trac}}(t) - F_{\text{brake}}(t) - F_{\text{res}}(v(t)) + F_{\text{slope}}(s(t)) \right)
\]  

(4.40)

According to (4.12) the force caused by inclinations of the track can be written as:

\[
F_{\text{slope}}(s) = \delta(s)mg.
\]

(4.41)
where $\delta(s)$ is the slope of the track, $m$ is the train’s mass, and $g$ is the acceleration of gravity. From this point, inclination $\delta(s)$ is considered constant for the inspected section and will be noted as $\delta$. Also $F_{trac}(t)$ and $F_{brake}(t)$ is combined and considered as constant, and will be noted as $F_{trac}$.

By assuming the second order form of traction resistance as:

$$F_{res}(v) = m \left( \alpha + \gamma v^2 \right), \quad (4.42)$$

the differential equation (4.40) is formed as:

$$\frac{dv(t)}{dt} = \frac{F_{trac}}{m} + g\delta - \left( \alpha + \gamma v(t)^2 \right)$$

(4.43)

Considering the initial condition for starting speed:

$$v(0) = v_0,$$

(4.44)

the solution of the equation can be formed as:

$$v(v_0, t, F_{trac}) = \frac{A}{m^\gamma} \tanh \left( \left( tA + m \arctanh \left( \frac{v_0 m^\gamma}{A} \right) \right) \right), \text{ where }$$

$$A = \sqrt{m^\gamma \left( F_{trac} + mg\delta - m\alpha \right)} \quad (4.45)$$

Another approach should be the linearization of the $F_{res}$ resistive force in the form of:

$$F_{res}(v) = m \left( \alpha + \beta v \right),$$

(4.47)

In this case the differential equation (4.40) is formed as:

$$\frac{F_{trac}}{m} + g\delta - (\alpha + \gamma v(t)) = \frac{dv(t)}{dt}$$

(4.48)

Solving the equation leads to the explicit form of the velocity function over time:

$$v(v_0, t, F_{trac}) = e^{-\beta t} \left( v_0 - \frac{F_{trac} + mg\delta - m\alpha}{m\beta} \right) + \frac{F_{trac} + mg\delta - m\alpha}{m\beta}$$

(4.49)

From this the time needed to reach speed $v_1$ from speed $v_0$ with a given tractive force $F_{trac}$ can be written as:

$$t(v_0, v_1, F_{trac}) = - \ln \left( \frac{(\delta g - \alpha - v_1 \beta) m + F_{trac}}{(-\beta v_0 + \delta g - \alpha) m + F_{trac}} \right)^{\beta^{-1}}$$

(4.50)

The distance driven can be obtained by integrating (4.49) from $t(0) = 0$ to $T$:

$$s(T) = \int_0^T v(t) dt$$

(4.51)

79
This way the distance driven with initial speed $v_0$, constant tractive force $F_{trac}$ in time $t$ is:

$$s(v_0, t, F_{trac}) = \frac{1}{m\beta^2} \left( (\text{-}\beta v_0 + \delta g - \alpha) m + F_{trac} \right) e^{-\beta t} + (v_0 + t g \delta - t \alpha) \beta - \delta g + \alpha) m + F_{trac} (\beta t - 1)$$  \hspace{1cm} (4.52)

From (4.53) the time needed to drive a given distance can be obtained as:

$$t(S) = \frac{1}{\beta ((\delta g - \alpha) m + F_{trac})} \left( ((\delta g - \alpha) m + F_{trac}) + (-\beta v_0 + \delta g - \alpha + S \beta^2) m + F_{trac} \right)$$

$$W \left( - ((\delta g - \alpha) m + F_{trac}) e^{\left(\frac{\beta v_0 - \delta g + \alpha - S \beta^2}{(\delta g - \alpha) m + F_{trac}}\right)} ((\delta g - \alpha) m + F_{trac})^{-1} \right)$$  \hspace{1cm} (4.53)

Where $W$ is one of the real branches ($W_0$ or $W_{-1}$) of the Lambert W function that provide feasible solution.

As an example the case the acceleration-coasting sequence is shown. The basics can be seen on Figure 4.22. The section is divided into two parts, the first is when the train accelerates at maximal tractive force from $v_0$ to $v_1$ at time $t_1$ and runs $s_1$ distance. This is the starting point of the coasting phase, which have $s_2$ distance.

Figure 4.22: The acceleration-coast problem
The main task in this problem is to find the acceleration time \( t_1 \) during which the train accelerates to \( v_1 \) so the overall travel time \( T \) is kept. The system of equations for the problem can be written as:

- The journey time \( T \) is fixed and is the sum of the travel times (\( t_1 \) and \( t_2 \)) of the two phases, see equation (4.54)
- The distance covered \( S \) is the sum of the distances of the two phases (\( s_1 \) and \( s_2 \)), see (4.55). For the calculation of distances (4.53) can be used.
- Finally, for the evaluation of (4.55) \( v_1 \) is needed which is dependent on \( t_1 \), and can be calculated based on (4.49).

\[
T = t_1 + t_2 \quad \text{(4.54)}
\]
\[
S = s_1 + s_2 = s(v_0, t_1, F_{\text{trac}}^{\max}) + s(v_1, t_2, 0) \quad \text{(4.55)}
\]
\[
v_1 = v(v_0, t_1, F_{\text{trac}}^{\max}) \quad \text{(4.56)}
\]

The solution of the set of equations (4.54)-(4.56) for \( t_1 \) is the following:

\[
t_1 = -W\left(\frac{(S\beta^2 + (-v_0 + T\alpha - Tg\delta)\beta + \delta g - \alpha)m - \beta Tf_{\text{trac}}}{e^{\beta T}F_{\text{trac}}^\beta}\right) + (4.57)
\]

By knowing \( t_1 \), the other variables can be easily calculated. When the speed limit restricts the utilization of the simple acceleration-coast strategy, a holding phase has to be calculated, as shown on Figure 4.23.

At this case the set of unknown variables are: \([v_1, v_2, t_1, t_2, t_3, s_1, s_2, s_3]\). Since speed holding occurs at the speed limit:

\[
v_1 = v_2 = v_{\text{max}} \quad \text{(4.58)}
\]

And therefore \( t_1 \) can be calculated from (4.50), since it’s the time needed to accelerate from \( v_0 \) to \( v_{\text{max}} \) with maximum tractive force, and \( s_1 \) is also determined:

\[
t_1 = t(v_0, v_{\text{max}}, F_{\text{trac}}^{\max}) \quad \text{(4.59)}
\]
\[
s_1 = s(v_0, t_1, F_{\text{trac}}^{\max}) \quad \text{(4.60)}
\]

Which leads to the following system of equations. The two last phases share the remaining journey time (4.61). And the remaining distance is also can be calculated as shown in (4.62).

\[
T - t_1 = T_r = t_2 + t_3 \quad \text{(4.61)}
\]
\[
S - s_1 = S_r = s_2 + s_3 = t_2v_1 + s(v_2, t_3, 0) \quad \text{(4.62)}
\]
Figure 4.23: The acceleration-hold-coast problem

The key point is to find the time of the speed holding phase, which can be calculated from (4.58)-(4.62):

\[ t_2 = -W \left( -v_2 \beta + \delta g - \alpha \right) e^{\frac{(-S_r + t_1 v_2) \beta^2 + v_2 \beta + \alpha - \delta g}{\beta^2 + (T_r \delta g - T_r \alpha + v_2) \beta + \alpha - \delta g}} (\delta g - \alpha)^{-1} + \right. 
\]

\[ \left. (\delta g - \alpha)^{-1} \left( (-S_r + t_1 v_2) \beta^2 + (T_r \delta g - T_r \alpha + v_2) \beta + \alpha - \delta g \right) \beta^{-1} \right) \]

Since the calculation of further phases are similar to the others shown above their interpretation is omitted in this section.

### 4.6.2.4 Optimization

The optimization problem is similar to the one described in Section 4.6.1.2, one has to find a speed profile with minimal energy consumption, that keeps the desired journey time. The generation of the speed profiles are based on a desired \( t_i (i = 1 \ldots n) \) journey times of each subsection. Though the train can not always satisfy the desired time, since the starting speed of the subsection could be too low for this task. Naturally one can provide a theoretical minimum for the \( t_i \) trip time from the speed limits, and subsection length:

\[ t_i^{\text{min}} = \frac{s_i}{v_{\text{reg},i}} \quad (4.64) \]

where \( t_i^{\text{min}} \) is the minimum allowed time, \( s_i \) is the length and \( v_{\text{reg},i} \) is the speed limit for the subsection. The overall journey time, that can be achieved with the given \( t_i \) vector
can only be calculated as the numerical evaluation of all sections is noted \( T^{calc}(t_i) \). The calculation of the energy consumption is similar to the previous algorithm (4.31) and is based on the generated speed profile and is noted as \( E(t_i) \).

The cost function of the optimum search is the weighted average of the normalized values of energy consumption and calculated journey time:

\[
J(t_i) = W_e \frac{E(t_i)}{E^{max}(t_i)} + W_t \frac{|T^{calc}(t_i) - T|}{\bar{T}},
\]

(4.65)

where \( W_e \) and \( W_t \) are weighting parameters, \( T \) is the desired journey time and \( E^{max}(t_i) \) is the energy consumption of the run with minimum feasible journey time.

The only restriction for the generated \( t_i \) vector is that the lower limit of the journey times are limited, while the keeping of the speed limits are evaluated implicitly in \( T^{calc}(t_i) \). The optimization problem can be formalized as:

\[
\begin{aligned}
\text{minimize} & \quad J(t_i) \\
\text{with respect to} & \quad t_i; i = 1, \ldots, n \\
\text{subject to} & \quad t_i^{min} \leq t_i
\end{aligned}
\]

(4.66)

4.6.2.5 Simulation results

The demonstration of the algorithm will consider the same section as the previous algorithm, with the same speed limits. Though in this case track inclinations are also considered, and therefore the division of the section is based on the speed limits and the track inclinations. This can be inspected on Figure 4.24.

Figure 4.24: Altitude diagram and speed limits of the section

Figure 4.25 shows the generated speed profiles for four different speed profiles. It is clearly visible - comparing with Figure 4.24 - how the algorithm utilizes the slopes.
Parallel the possibilities provided by the longer journey allow to use the more coasting, which can be deduced by the gradient of the speed-distance function.

Finally the blue line in Figure 4.25 and Figure 4.26 is a typical example of the tight timetable. There is not any available spare time, thus the maximal allowed speed has to be utilized on each section, coasting can not be used and the slopes can not be exploited. The result is spectacularly presented by the largest energy consumption growth.

### 4.7 Driver Advisory Algorithm

#### 4.7.1 Model Predictive Optimization Problem

In order to achieve the aim of the optimization criterion i.e. to reduce energy consumption while keeping time we have two possibilities: first is to design the speed profile up
to the end of the section continuously, which obviously requires excessive computational effort. The second option is to use a model predictive control algorithm [Grüne and Pannek, 2011] that considers the characteristics of the track in some distance ahead. This approach allows to take advantage of the slopes of the track, although a reference run is needed to ensure the time keeping property of the algorithm. The first step of the algorithm development is selecting the control inputs of the model and the basis of discretization.

Such systems generally use the desired acceleration or the tractive/braking force for model input, while in this case the desired speed has been chosen as input. Both approaches have their strengths and weaknesses. While choosing the forces or acceleration demand as inputs, (4.5) can be easily handled as input constraint while the satisfaction of the speed limits in (4.6) is a more complicated task. Choosing the speed profile as input makes (4.6) a simple bound on input and the handling of the force constraints is much more feasible. Regarding the base of discretization a space-discretized approach has been chosen instead of the classic time-based discretization.

The modelling paradigms used in the construction of the system are summarized by Fig. 4.27. The space is discretized with sampling points \( s_i \). The method does not require equidistant steps \( \Delta s_i = s_i - s_{i-1} \neq const \) though in the simulation results presented later in the chapter, equidistant scaling is used. Speed limits \( v_{lim}(s_i) \) are given as a function of position. The prerequisite of the optimization is a known reference run \( (v_{ref,i}, t_{ref,i}) \). The altitude \( h_i \) of the track is sampled and considered to be known at all \( n \) points of the prediction horizon, and modelled as a first order hold condition assuming constant steepness between two samples.

The optimization problem is to find the suitable sequence of velocities \( v_i (i = 1..n) \)
while the constraints of the original problem given in (4.5) and (4.6) must still be satisfied. The conditions given so far do not guarantee the desired journey time, for this task the following terminal constraints are introduced concerning the reference run:

\[ v_n = v_{ref,n} \quad (4.67) \]
\[ t_n = t_{ref,n} \quad (4.68) \]

Simultaneously the task is to minimize the energy consumed by the tractive effort:

\[ E_n^0 = \sum_{i=1}^{n} F_{trac,i} \Delta s_i \rightarrow \min \quad (4.69) \]

The introduction of constraint (4.67) means that the speed of the input sequence \((v_i)\) and the reference run must be equal at the end of the prediction horizon, ensuring that energy saving of the optimal solution is not achieved by the loss of motion energy. It is assumed that the reference run fulfils (4.7) thus from every point of the run, the journey time is tenable. The terminal constraint on time (4.68) ensures this feature on the optimized run as well.

The conditions (4.68) and (4.69) are in contradiction since achieving minimal energy consumption can be mostly achieved at lower speed, which leads to increasing journey time. Equation (4.68) is written as a terminal condition but can also be considered as a condition that tightens the set of feasible solutions, since travel time \(t_n\) at the end of the horizon can be explicited with \(t_0\) and the \(v_i\) sequence with the well-known kinematic relationship:

\[ t_n = t_0 + \sum_{i=1}^{n} \frac{2\Delta s_i}{v_{i-1} + v_i} \quad (4.70) \]

Furthermore (4.68) can not always be satisfied, since when \(t_0 > t_{ref,0}\), the train reaches the current step in delay; it is possible that speed limit \(v_{lim}\) does not permit the generation of a sequence \(v_i\) that fulfils the condition. Therefore it may be expedient to rewrite the condition to the minimization of the time difference instead of keeping the exact run time changing it to a terminal cost function:

\[ |t_n - t_{ref,n}| \rightarrow \min \quad (4.71) \]

The problem in this form results in a multi-criteria optimization problem. To combine the two conditions weighted sums can be used:

\[ J(v_i) = W_E E_n^0 + W_t |t_n - t_{ref,n}| t_{ref,n} - t_{ref,0}, \quad (4.72) \]

where \(J(v_i)\) is the single criterion objective function of the problem where \(W_E \in R^+\) is the weight of the energy consumption costs and \(W_t \in R^+\) is the weight of the delay costs. Choosing the correct weighting highly affects the output of the algorithm as it will be shown later.
Though the calculation of $E_n^0$ is given in (4.69) as the sum of work in each section, it is simpler to calculate the change of the energy of the system in the $i$th step as the function of the energy loss from resistance, and the change in the potential and kinetic energy:

$$\Delta E_i = \frac{1}{2}m(v_i^2 - v_{i-1}^2) + \Delta h_i mg + F_{res,i}\Delta s_i$$

(4.73)

In the case where $\Delta E_i$ is positive, the traction engine consumes energy $E_{trac,i}$, otherwise it is braking. Since the discussed model does not assume regenerative braking $E_n^0$ only considers the positive part of $\Delta E_i$:

$$E_{trac,i} = \max(0, \Delta E_i)$$

$$E_n^0 = \sum_{i=1}^{n} E_{trac,i}$$

(4.74)

### 4.7.2 Optimization algorithm

The function to be minimized (4.72) is non-linear but continuous for $\forall v_i \in R^+$. The feasible set has a boundary condition derived from the speed limit:

$$0 \leq v_i \leq v_{lim,i}, \forall i \in (0..n-1)$$

(4.75)

Another constraint may be the limitation of the difference of neighbouring elements of the $v_i$ vector to keep the maximum traction and braking forces:

$$-F_{\text{max brake}}(v_{i-1}, v_i) \leq \frac{\Delta E_i}{\Delta s} \leq F_{\text{max trac}}(v_{i-1}, v_i)$$

(4.76)

There are three ways of handling the constraint exposed in (4.76).

- The first and possible the most obvious solution is that constraint should be implemented into the generation of feasible set of the optimization algorithm. However, this method has high computational requirements, which slows down the optimization.

- The second method should be the utilization of the constraint during the evaluation saturating the actual change in speed. Although this method is much faster than the first one, it leads to a loss of information from the $v_i$ vector and generates zero-gradient areas of the objective function making it difficult for the algorithm to find the optimum.

- The third possibility is omitting (4.76) through the optimization, and only take it into account at the application of the first step of the solution.

Since the second possibility does not always lead to the optimal solution and the first one makes the optimization problem too complex numerically, the third solution is used for the algorithm.

The main steps of the algorithm are as follows (see Fig.4.28.):
• **Step 1**: Acquire signal $t_i$ and measure the state variable $v_i$ of the system.

• **Step 2**: Minimize $J(v_i)$ (4.72) with respect to (4.67) and (4.75). As a result the optimal predicted speed profile is given:

$$\hat{v}_i = [v_{i+1}, v_{i+2}, \ldots, v_{i+n}]^T$$ (4.77)

• **Step 3**: The feedback value for the control loop is defined as the first element ($v_{i+1}$) of the solution vector (4.77) gained from the Nonlinear Model Predictive Control (NMPC) controller.

• **Step 4**: The actuator checks and validates the given speed command with respect to (4.76) and saturates the force command if needed.

**Figure 4.28**: The NMPC control loop

Regarding the optimization algorithm the handling of non-linear objective function and boundaries should be fast. For the optimization task, sequential quadratic programming is used.

**4.7.3 Remarks**

Note that the selection of $\Delta s_i$ and $n$ is based on the track. For example, in the case of flat track with few speed limits, it is enough to use relatively small $n$. However, in the case of undulating tracks with frequent speed limits, large $n$ is needed.

**4.8 Simulation results**

The developed predictive optimization method is tested by simulation on a real railway section. The simulation is based on the test runs made by SBB between Fribourg and
Figure 4.29: Demonstration of the energy-efficient method. Speed limits and the altitude diagram of the section.

Bern with an approximate length of 30 km, see [Meyer et al., 2007]. Although the test of the SBB used regenerative braking the results are comparable.

Figure 4.29(a) shows the speed limits of the track and a possible reference run that does not take advantage of the inclinations of the track and takes 19 minutes and 20 seconds to cover the distance. The purpose of the simulation is to show that by using predictive optimization, energy can be saved while keeping the same travel time between the two stations. The simulation used the theoretical tractive effort curve of the locomotive RE460 and a total train mass of 300t. To run the algorithm the altitude diagram is needed, which is shown in Figure 4.29(b). The simulation is carried out with $\Delta s = 100m$ equidistant spacing and $n = 50$ long prediction sections resulting in 5 km prediction distance. This means that in every decision step, the algorithm takes the inclinations of the next 5 km into consideration.

Figure 4.30(a) shows the distance-speed diagram of the results of the algorithm compared to the limits and the reference run. Naturally, owing to the continuous time keeping property of the method, the optimized curve does not significantly deviate from the reference run. Figure 4.30(b) shows that at the end of the run the final delay is approximately 1 sec and the algorithm keeps the delay in $[-6; 8]$ sec during the entire run. Due to the nature of the algorithm delays may occur on the open track but they are eliminated by the time of arrival point. Figure 4.30(c) shows the cumulative energy consumption of both the optimized and the reference run. Examining the figures together one can find the key points of energy saving. The optimization in this scenario resulted in approximately 10% energy saving reducing the consumption from 141.7 kWh to 127.3 kWh.
4.9 Sensitivity and Robustness

When designing an NMPC-like system one should always examine the obtained control loop regarding robustness and sensitivity since the real system never exactly coincides with its mathematical model. This means that in practice the behaviour of the real system will deviate from its mathematical model. The solution of the optimization contains weighted average for the terminal costs and the tuning of these weights is also an interesting point that affects the performance of the method. This section shows two case studies: one for parameter uncertainty and another for the tuning of the optimization weights.

4.9.1 Robustness to Parameter Uncertainty

The deviations of the model from the real plant may come from several sources, such as modelling errors neglecting or simplifying behaviours, parameter uncertainty, measure-
ment or actuation differences or disturbances acting on the system. Furthermore, the numerical integration method applied at the discretization of the continuous model for solving the finite horizon optimal control problem results in the fact that the predicted trajectory does not match the trajectory of the plant.

The proposed optimization method is not exempt from these problems. Throughout the development of the model some resistance forces such as curve resistances are ignored, and first order hold method is chosen for the discretization of the speed profile as well as for the inclinations of the track assuming linear steepness between two sampling points for the optimization algorithm. External disturbances, such as the effect of the wind on the resistance forces can not be taken into consideration. The condition of the rail and the wheels, weather effects such as humidity, temperature, snow or even the dirt on the track highly influence the interaction of the train and the track through the loss of the achievable adhesion force resulting in the reduced acceleration and deceleration rates. Though the wind and the adhesion conditions could be estimated the solution will always contain some uncertainty.

Two parameters are chosen to examine robustness. One is the mass of the train and the other is the inaccurate modelling of the resistance forces.

![Graphs](attachment://graphs.png)

- (a) Speed profile
- (b) Time delay during the journey
- (c) Cumulative energy consumption

Figure 4.31: Analysis of the uncertainty of the train mass with different assumed weight ratio ($m_{\text{real}}/m_{\text{optim}}$)

91
Figures 4.31(a), 4.31(b) and 4.31(c) show the simulation results where the mass parameter for the optimization deviates from the half to the double of the real train’s mass. Normally the train’s mass ought to be known with the maximum relative error of 30-40% so these scenarios can be considered extremist. After the inspection of the runs it is shown that the closed loop ensures the keeping of the terminal constraints. The underestimation of the weight may lead to the loss of the time keeping property, and the overestimation can lead to unnecessary overconsumption, even though these phenomena can not be directly inspected on this scenario but it can be deduced from the investigation of the in run delays and energy consumption diagrams.

Regarding the uncertainty of the resistance force $F_{\text{res}}$, Figure 4.32 shows that the model is robust to the proportional error of the resistance force.

![Figure 4.32: Speed profile with different considered $F_{\text{res}}$ values where the assumed forces proportionally deviating from 0.5 to 2.0](image)

4.9.2 Optimization Weights

Another interesting point is the appropriate selection of the $[W_t, W_e]$ weights of the cost function (4.72) and how sensitive the optimization algorithm is to the change in these weights. As previously mentioned the terminal constraint for keeping the time (4.68) is changed to a terminal cost function (4.71) to ease the solving of the finite horizon optimization problem. Though the need to keep time at the end of the horizon still remains so one must choose the weights of the cost function to ensure this behaviour. For the results presented in the simulation section $(W_t/W_e = 100)$ is chosen to overweight the time keeping at the expense of energy consumption. To examine the effects of the weights several test simulations are performed on the model under the assumption that all parameters are known with high certainty.

Figures 4.33(a), 4.33(b) and 4.33(c) show the simulation results where $(W_t/W_e)$ are changed from 1000 to 10. After the inspection of results the most important conclusion is that by lowering the proportion of the weights the time keeping constraint is violated. Around $(W_t/W_e \approx 45)$ the time keeping is lost during the run, but restored at the end at the cost of increased energy consumption. With values below 30 an energy consumption gain can be achieved but at the cost of unacceptable delay.
4.10 New results

4.10.1 Thesis III

I have developed a parameter identification method for the determination of the coefficients of the longitudinal propulsion resistance formula of the trains, which uses the track data, energy consumption, speed and position information of the telemonitoring system. The approach is model based meaning that the searched parameters are used to evaluate the energy consumption of the train, which is compared to the measured energy consumption leading to a minimization problem.

Related publications: [Aradi et al., 2015a]

- I have studied the state of the art of the modelling techniques of simplified longitudinal propulsion resistance and also the methodologies of determining propulsion resistance through measurement.

- I have defined the formula of the propulsion resistance, which - as a simplified model - considers the locomotive’s mass and the number and mass of the rolling
stock:

\[ F_{\text{loc}}^{\text{res}}(v) = m_{\text{loc}} \ast (\alpha_l + \beta_l v) + \gamma_l v^2 \]  
\[ F_{\text{stock}}^{\text{res}}(v) = m_{\text{stock}} \ast (\alpha_s + \beta_s v) + n_{\text{stock}} \ast \gamma_s v^2 \]  

(4.78) (4.79)

- I have used the resistance formula for elaborating the energy consumption model of the train.

- The minimization of the difference between the measured and calculated energy consumption leads to the determination of the searched parameters:

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{with respect to} & \quad x = [\alpha_l, \beta_l, \gamma_l, \alpha_s, \beta_s, \gamma_s, \xi_j, \psi_j]; j = 1, \ldots, k \\
\text{subject to} & \quad 0.9 \leq \xi_j \leq 1.1 \\
& \quad 0.0 \leq \psi_j \leq 1.0 \\
& \quad x_i \geq 0.0; i = 1, \ldots, 6
\end{align*}
\]  

(4.80)

- The developed methodology was tested on actual measurements (Figure 4.11), by using the data of the electric locomotive series 431 of the Hungarian State Railways.

### 4.10.2 Thesis IV

I have designed a method for the determination of the energy-optimal speed profile for trains. The proposed algorithm considers the parameters of the train, the speed limits of the given section and the inclinations of the track to generate a time-keeping yet energy efficient speed trajectory between two stations.

Related publications: [Aradi et al., 2013b],[Aradi et al., 2014a]

- I have shown the effects of the inclinations of the track on the energy consumption of the railway operation, and explicitly defined a mathematical formula of the time and energy consumption of driving strategies for different special conditions.

- By using these formulas and by dividing the inspected section into subsections with constant speed limits and approximately constant track gradients, the speed profile can be determined as a function of the required journey time.

- Hence, the optimal speed profile can be generated as a function of the journey time vector of the subsections. The weighted average of the cumulated running time and energy consumption could provide a possible objective function for the determination of this vector.

\[ J(t_i) = W_e \frac{E(t_i)}{E_{\text{max}}(t_i)} + W_t \frac{T_{\text{calc}}(t_i) - \bar{J}}{\bar{J}} \]  

(4.81)

- I have validated the functionality and performance of the algorithm by using the inclination and speed limit parameters of a real railway section (Figure 4.25).
4.10.3 Thesis V

I have designed a driver advisory algorithm, which - based on a previously calculated or recorded reference run - is able to provide information about the optimal speed choice ensuring minimal energy consumption while keeping the journey time and the speed limits. The method determines the traction force requirements based on the reference run and the current state of the train.

Related publications: [Aradi et al., 2013b], [Bécsi et al., 2013], [Aradi et al., 2014a]

- The model uses a predictive approach with the space-discretization of the section considered in the prediction, defining a speed value for each step considering the speed limits and other restrictions.

- Train driving is modelled as a control loop that consists of the measurement of the train’s state, the optimum search at the prediction interval and actuation (Figure 4.28).

- The keeping of the journey time is ensured by considering the speed \( v_n \) of the reference run at the end of the prediction horizon as a terminal constraint and also by minimizing the delay calculated at the same position.
Chapter 5

Conclusions and Future Research

5.1 Practical Aspects

The Vehicle-to-Infrastructure networks and vehicle telemonitoring systems have been spreading rapidly in recent years. The functionalities of the existing systems are expanding which results in the growth of the data volume generated by these applications. The field of rail traffic is not an exception to this trend. The handling of this data volume is simultaneously a threat and an opportunity. An other possible direction of development is the automated train operation which is a known technology in some areas of fixed track applications such as subways. The data gained from the telemonitoring system can be used for extending these functionalities with the aspects of energy efficiency. Moreover this optimization may be raised to line and network level.

The problems generalized above require many low-level solutions and my research is aimed at finding the solutions to some of them, keeping in mind the practical applicability.

The achievements of Thesis 1 (Server Solutions for On-line Telemonitoring Systems) can be utilized at the design of railway telemonitoring systems that receive and process data generated on the on-board units of a train and also interact with them.

The functionalities of the telemonitoring system can be extended by developing an 'intra-train' network on trains which do not have a dedicated communication line for miscellaneous data. Some answers to this problem are given by Thesis 2 (Development of Vehicle On-board Communication System for Rough Environment).

With the help of the algorithm outlined in Thesis 3 (Parameter Estimation), based on the data of the telemonitoring system, the parameters of the longitudinal resistance force equations can be evaluated without conducting special measurements. These results can be used in energy-optimization algorithms and also in other calculations of railway dynamics.

Thesis 4 (Development of the Optimization Algorithm) provides an algorithm for determining the energy efficient speed profile of a train considering track inclinations and speed limits. On the hand this will give a reference for the algorithm designed in the next thesis and also can be used for the evaluation and training of engine drivers.
Last, the algorithm given in Thesis 5 (Driver Advisory Algorithm) can be a basis of an on-board advisory system that helps the train driver to keep the journey time and save energy simultaneously providing real-time information and recommendations. Also this algorithm can be a basis for automated and energy-efficient train operation.

5.2 Future research

Besides the discussion of the questions of designing and operating telemonitoring system in the field of transportation, the dissertation points out that the possibilities of such systems make them much more than a simple data transfer and storage application, but rather with the appropriate processing of data they could become the fundamental tool for the optimized operation of the network in the hands of operators.

Vehicle trajectories and speed can be optimized by using the tools of numeric optimization and control theory with the result of a significant gain in energy consumption and environmental impact. This can be particularly true in the case of railways, where due to the large mass of trains – substantial energy savings can be achieved with relatively low investments and operating costs of telematics.

I have developed an algorithm for a locomotive on-board driver advisory system that is capable of calculating a combined energy-optimal and journey time keeping speed profile between two stations based on the data provided by the telemonitoring system. Based on these achievements, recommendations can be given to the train driver to adjust optimal speed or for the direct handling of actuators. Further research should mainly focus on this problem group.

Although it is not linked closely to the optimization theoretic approaches of the dissertation, one possible research direction should be the presentation of the recommendations (e.g. recommended speed or recommended engine control states) towards the train driver, which needs to consider human factors and must meet the requirements of accurate, efficient but most importantly safe transportation. One of the main questions concerning the different driver advisory systems is the adequate visualization and timing of the information provided to the driver, such as the contents of the display, the rate of refreshing, ergonomics, and the ease of understanding data. All these aspects not only affect efficiency but also traffic safety and therefore their research is an important task.

Another research direction could be the extension of the optimization to railway line or even network level by using the recent achievements of the field of transportation control. The algorithms presented in the dissertation focus on the control of a single train, where the traffic conditions are only considered as external conditions. The next step could be the optimal control of a railway line, since the telematics system is capable of providing the position and speed of all trains in almost real-time. Hence, the optimization could take into account the following, followed or crossing trains and the number of unnecessary brakings can be further reduced. Next research should be conducted for the optimization of the rail network of whole regions or even countries. Though at the field of road traffic modelling this is the level of macroscopic approaches, in the field of
railways, the relatively low number of trains enables the handling of this problem on a train level, or with some mesoscopic approach.

As can be seen from the above mentioned examples, the field of researching optimal control strategies of railway operations holds great opportunities which - in accordance with the current efforts of the developed countries - may facilitate the reduction of energy consumption and therefore the $CO_2$ emission of transportation.
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>APN</td>
<td>Access Point Name</td>
</tr>
<tr>
<td>ATO</td>
<td>Automatic Train Operation</td>
</tr>
<tr>
<td>CAN</td>
<td>Controller Area Network</td>
</tr>
<tr>
<td>CPU</td>
<td>Central Processing Unit</td>
</tr>
<tr>
<td>EIA</td>
<td>Electronic Industries Alliance</td>
</tr>
<tr>
<td>ELS</td>
<td>Electronic Logbook System</td>
</tr>
<tr>
<td>EMC</td>
<td>Electromagnetic Compatibility</td>
</tr>
<tr>
<td>ERP</td>
<td>Enterprise Resource Planning</td>
</tr>
<tr>
<td>ERTMS</td>
<td>European Rail Traffic Management System</td>
</tr>
<tr>
<td>ETCS</td>
<td>European Train Control System</td>
</tr>
<tr>
<td>FIFO</td>
<td>First In First Out</td>
</tr>
<tr>
<td>FMS</td>
<td>Fleet Management System</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>GPRS</td>
<td>General Packet Radio Service</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobilecommunication</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>ISO</td>
<td>International Organization for Standardization</td>
</tr>
<tr>
<td>ISP</td>
<td>Internet Service Provider</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>LPV</td>
<td>Linear Parametric Varying</td>
</tr>
<tr>
<td>MMI</td>
<td>Man-Machine Interface</td>
</tr>
<tr>
<td>NMPC</td>
<td>Non-linear Model Predictive Control</td>
</tr>
<tr>
<td>OBU</td>
<td>On-Board Unit</td>
</tr>
<tr>
<td>OSI</td>
<td>Open Systems Interconnection</td>
</tr>
<tr>
<td>PI</td>
<td>Proportional Integral</td>
</tr>
<tr>
<td>SMS</td>
<td>Short Message Service</td>
</tr>
</tbody>
</table>
**SQP** – Sequential Quadratic Programming

**TCP** – Transmission Control Protocol

**TEM** – Transverse Electromagnetic Wave

**TIA** – Telecommunications Industry Association

**TTPS** – Traction Technology Planning System

**UDP** – Universal Datagram Protocol

**UIC** – International Union of Railways

**UID** – Unique Identifier

**UTC** – Universal Time Coordinated

**UTF** – Unicode Transformation Format

**UTP** – Unshielded Twisted Pair

**VPN** – Virtual Private Network

**WGS84** – World Geodetic System ’84

**XML** – Extensible Markup Language

**XSD** – XML Schema Definition
List of Figures

2.1 Positioning the telemonitoring system in the global informatics structure of MÁV ................. 8
2.2 System architecture .............................................................. 9
2.3 The MMI of the on-board unit ............................................... 10
2.4 M/G/1 queue model of the telemonitoring service with one server ...... 15
2.5 Control system overview ......................................................... 23
2.6 Control synthesis ................................................................. 24
2.7 Poles of the first control scheme as a function of $\Delta_M$ .................. 25
2.8 Poles of the second control scheme as a function of $\Delta_M$ ............... 26
2.9 Simulated service time and incoming request for control validation ...... 27
2.10 Queue length and control schemes used for the actual state ............... 27
2.11 Estimation of $c_{serve}$ ........................................................... 28
2.12 The control input ($x_{serve}$) and the resulting processor load (cycle time) . 28

3.1 Train network structure .......................................................... 32
3.2 UIC 558 Series Cable and Connector ......................................... 32
3.3 General transmission line [Hui Tat, 2011] ...................................... 33
3.4 Transmission line section [York A., 2013] .................................... 34
3.5 Terminated transmission line circuit [Hui Tat, 2011] ............................ 35
3.6 EIA-485 network structure ....................................................... 36
3.7 CAN bus structure [ISO, 2003] ................................................. 38
3.8 Laboratory test, achievable CAN speed with different termination and length .......................................................... 40
3.9 Proposed network topology ....................................................... 41
3.10 Network state diagram ............................................................ 42
3.11 Network look-up example ........................................................ 43

4.1 Energy consumption compared to gross mass on Line 80 .................... 51
4.2 The energy consumption differences of the inspected journeys ............. 52
4.3 The energy consumption differences of the individual engine drivers .... 53
4.4 Map of Line 80 between Miskolc and Nyékládháza ............................ 54
4.5 Possible indicators of energy consumption ...................................... 55
4.6 The effect of braking on relative energy consumption .......................... 55
List of Figures

4.7 Forces acting on the train .............................................. 57
4.8 Comparison of propulsion resistance equations [Iwnicki, 2006] ........... 59
4.9 Line 1. of the Hungarian State Railways. Section between Budapest-
Kelenföld and Tatabánya ................................................ 61
4.10 Track elevation, speed trajectory, tractive force and energy consumption
of sample runs used for parameter estimation ............................. 63
4.11 Measured and calculated (with the found parameters) tractive forces of
Train No.1. and 2. ............................................................. 66
4.12 Measured and calculated cumulative energy consumption of Train No.4.
and 5. .............................................................................. 67
4.13 Bounds of the altered altitude diagram, and the identified speed resistance
diagrams generated with the parameters of Train No.1 .................. 68
4.14 Bounds of the altered energy consumption diagram of Train No.1, and
the identified speed resistance diagrams generated with the parameters of
Train No.1 ........................................................................ 68
4.15 Resistance diagrams of Train No. 1 with asynchronous data recordings ... 69
4.16 Example division of a given section ........................................ 70
4.17 Example section. Speed limit and speed profile with the lowest journey
time ................................................................................. 74
4.18 Generated speed profiles with multiple journey times ...................... 75
4.19 Energy consumption of the generated speed profiles with multiple journey
times ................................................................................. 76
4.20 Energy consumption of the generated speed profiles with multiple journey
times. (Track inclinations, Speed and Forces) .............................. 77
4.21 Possible optimal strategies under different conditions of track gradient
and journey time .................................................................. 78
4.22 The acceleration-coast problem ............................................. 80
4.23 The acceleration-hold-coast problem ....................................... 82
4.24 Altitude diagram and speed limits of the section ........................... 83
4.25 The generated speed profiles with multiple journey times ............... 84
4.26 Energy consumption of the generated speed profiles with multiple journey
times ................................................................................. 84
4.27 The predictive optimization problem ....................................... 85
4.28 The NMPC control loop .................................................... 88
4.29 Demonstration of the energy-efficient method. Speed limits and the alti-
tude diagram of the section. .................................................. 89
4.30 Demonstration of the energy-efficient method of the reference run (dashed)
and the output of the algorithm (solid) ....................................... 90
4.31 Analysis of the uncertainty of the train mass with different assumed weight
ratio ($m_{real}/m_{optm}$) .......................................................... 91
4.32 Speed profile with different considered $F_{res}$ values where the assumed
forces proportionally deviating from 0.5 to 2.0 ......................... 92
4.33 Analysis of the ratio of the weighting factors $W_l/W_e$ ..................... 93

102
# List of Tables

2.1 The 7-layer model of the communication subsystem ........................................ 12  
3.1 Function allocation of the UIC 558 cables .................................................. 33  
3.2 Laboratory test results ................................................................................... 39  
3.3 Example with 9 unit train from the view of the left side end of the network 44  
3.4 Determination of orientation of each unit ...................................................... 44  
3.5 Field measurement, CAN network with different termination, using UIC 558 interconnect cable .......................................................... 46  
4.1 Stations at the studied section of Line 80. ...................................................... 54  
4.2 Empirical Formulas for Propulsion Resistance, Freight Rollingstock [Iwnicki, 2006] ................................................................. 58  
4.3 Main properties of the 0431(V43) locomotive ............................................... 61  
4.4 Basic parameters of the example trains ......................................................... 62  
4.5 Results of the error analysis. The maximum of the absolute and relative error of the generated running resistance diagrams at different speeds (Train No. 1) ................................................................. 69  
4.6 The possible speed profile patterns used for the method ............................... 71  
4.7 Kinematic equations of the different phases ............................................... 72
References


