

**The University of Nottingham**  
**Faculty of Engineering**  
**School of Electrical and Electronics Engineering**



**Effective Text Compression For Short Messages**

AUTHOR : Lau Yee Kuan  
SUPERVISOR : Dr. Lim Wee Gin  
DATE : April 20, 2007

Fourth year project thesis submitted in partial fulfillment of the requirements of the degree of **Master of Engineering**

## **Abstract**

Recently, the scope of the usage of SMS (Short Message Service) is beyond just text messaging. SMS is prevalently used for Business-to-Business communication, as well as Machine-to-Machine communication. The widespread use of SMS (Short Message Service) as an effective business communication tool projects the need to enhance the main constraint of SMS. Its major drawback of having a maximum number of 160 characters for one SMS is not sufficient for most business purposes, such as alert notifications and business reports.

Hence, this thesis explores new schemes to effectively compress SMS and subsequently reducing the amount of SMS that was required to be sent. Depending on the arbitrary SMS typed, the compression ratio can improve significantly. With this, enterprises will be able to increase profit through cost minimization. Following that, the SMS will be sent using a modem from the Personal Computer (PC). Ultimately, the whole process of compression up till the sending SMS was implemented in a Visual Basic interface.

## **Acknowledgement**

I would like to express my heartfelt appreciation to my supervisor Dr. Lim Wee Gin for his guidance and assistance to complete this project. Extended thanks to last year's University of Nottingham Malaysia Campus Final Year Project student, Gan Jia Jian under the supervision of Dr. Lim Wee Gin, who laid down the thesis for this project.

Last but not least, I would like to thank my family members who provided constant support throughout the duration of the project. A big thank you goes out to everyone who has contributed, be it major or minor, to our project.

## Table of Contents

<b>Chapter 1: Introduction.....</b>	<b>1</b>
1.0 Overview .....	1
1.1 Why SMS?.....	1
1.2 Data Compression .....	2
1.2.1 Huffman Coding.....	3
1.3 SMS .....	4
1.4 Sending SMS through GSM modem .....	5
1.5 Project Outline.....	5
1.6 Proposed Deliverables.....	6
1.7 Thesis Structure.....	6
<b>Chapter 2: Literature Review .....</b>	<b>7</b>
2.0 Overview .....	7
2.1 Part I: Communication.....	7
2.1.1 Data Compression.....	7
2.1.2 Huffman coding .....	8
2.1.2.1 Generating Huffman Tree.....	9
2.2 Part II: Communication.....	12
2.2.1 SMS .....	12
2.2.1.1 AT Commands.....	12
2.2.1.2 Protocol Data Unit (PDU) .....	13
2.2.1.3 Parameter Descriptions .....	14
2.2.1.3.1 SCA (Service Center Address) .....	14
2.2.1.3.2 Protocol Data Unit Type (First Octet).....	14
2.2.1.3.3 Message Reference Number (TP-MR) .....	14
2.2.1.3.4 Destination Address/Originator Address.....	14
2.2.1.3.5 Protocol Identifier (TP-PID).....	15
2.2.1.3.6 Data Coding Scheme (TP-DCS) .....	15
2.2.1.3.7 Service Center Time Stamp (STCS) .....	15
2.2.1.3.8 Validity Period (TP-VP) .....	15
2.2.1.3.9 User Data Length (TP-UDL) and User Data (TP-UD) .....	16
<b>Chapter 3: Methodology .....</b>	<b>17</b>
3.0 Overview .....	17
3.1 The Main window .....	17
3.2 Part I: Compression .....	18
3.2.1 (A) Frequency Table and Manual Assignment of Alphabets .....	18
3.2.2 (B) Check for a specific pattern of the original SMS .....	23
3.2.2.1 Pattern detection of the arbitrary SMS.....	23
3.2.2.2 Further reduction of alphabet set size.....	24
3.2.3 Compression – Huffman coding .....	25

3.3	Part II: Communication.....	28
3.3.1	Sending SMS – Through Hyper Terminal .....	28
3.3.1.1	Step 1: Hyper Terminal - Send Text SMS.....	28
3.3.1.2	Step 2a: Hyper Terminal - Send PDU SMS (Single SMS) .....	29
3.3.1.3	Step 2b: Hyper Terminal - Send PDU SMS (Concatenated SMS).....	31
3.3.2	Step 3: Visual Basic – Send PDU SMS (Single/Concatenated) .....	34
3.3.3	Receiving SMS – PDU mode .....	38
3.3.3.1	Receiving Single SMS .....	38
3.3.3.2	Receiving Concatenated SMS .....	39
<b>Chapter 4:</b>	<b>Results and Discussion .....</b>	<b>40</b>
4.0	Overview .....	40
4.1	Part I: Compression .....	40
4.1.1	Relationship between Compression Ratio and Amount of SMS.....	40
4.1.2	Discussion of Compression Ratio for Single SMS .....	42
4.1.3	Discussion of Compression Ratio for Concatenated SMS.....	43
4.1.3.1	Case A: Good SMS example .....	44
4.1.3.2	Case B: Average SMS example.....	46
4.1.3.3	Case C: Concatenated SMS – Bad SMS example .....	48
4.2	Verification of the Receipt of SMS.....	50
4.2.1	Receiving Single SMS .....	50
4.2.2	Receiving Concatenated SMS.....	51
4.3	Summary .....	52
<b>Chapter 5:</b>	<b>Conclusion .....</b>	<b>53</b>
5.1	Part I: Compression .....	53
5.2	Part II: Communication.....	53
5.3	Future Development.....	53
<b>Reference.....</b>		<b>54</b>
<b>Appendix A.....</b>		<b>56</b>
<b>Appendix B.....</b>		<b>57</b>

## List of Tables

Table 2.1 : Steps to generate Huffman Tree and codebook .....	11
Table 2.2 : List of commonly used AT commands .....	12
Table 2.3 : SCA format .....	14
Table 2.4 : Destination Address or Originator Address .....	14
Table 3.1 : The first step of the effective compression method used .....	19
Table 3.2 : The second step of the effective compression method used .....	23
Table 3.3 : Pattern detection of the arbitrary SMS and the size of the respective alphabet sets .....	23
Table 3.4 : Step-by-step description of the Compression in Send SMS window .....	27
Table 3.5 : Different methods to send SMS .....	28
Table 3.6 : AT commands used to send SMS in Text Mode .....	29
Table 3.7 : AT commands used to send Single SMS in PDU Mode .....	30
Table 3.8 : First part of the Concatenated SMS - Explanation of PDU string .....	32
Table 3.9 : Second part of the Concatenated SMS - Explanation of PDU string .....	33
Table 3.10: Step-by-step description of the Communication in Send SMS window ..	35
Table 3.11: The Sending Status window for Single SMS .....	36
Table 3.12: The Sending Status window for Concatenated SMS .....	37
Table 3.13: Describes the SMSC information .....	38
Table 3.14: Length of PDU string (in octets) is obtained starting from first octet onwards .....	38
Table 3.15: User-Data length, User-Data header and User-Data (arbitrary SMS)....	38
Table 3.16: Describes the SMSC information .....	39
Table 3.17: Length of PDU string (in octets) is obtained starting from first octet onwards .....	39
Table 3.18: User-Data length, User-Data header and User-Data (arbitrary SMS)....	39
Table 4.1 : Two main concepts implemented in the project.....	40
Table 4.2 : Size per one SMS for Single SMS and Concatenated SMS .....	41
Table 4.3 : Single SMS analysis .....	41
Table 4.4 : Concatenated SMS analysis .....	41
Table 4.5 : Single SMS - Compression Ratio of various scenarios .....	42
Table 4.6 : Concatenated SMS – Good example SMS .....	44
Table 4.7 : Concatenated SMS – Average example SMS .....	46
Table 4.8 : Concatenated SMS – Bad SMS example .....	48
Table 4.9 : Verification of results – Single SMS example .....	51
Table 4.10: Verification of results – Concatenated SMS example .....	52

## List of Figures

Figure 1.1: The whole Data Compression process = Modeling + Coding .....	3
Figure 1.2: Two major families of Data Compression .....	3
Figure 1.3: Wavecom Fastrack modem M1206B .....	5
Figure 1.4: Block Diagram - Project Outline .....	5
Figure 2.1: ASCII character set .....	7
Figure 2.2: Pseudocode to generate Huffman tree .....	9
Figure 2.3: SMS SUBMIT format .....	13
Figure 2.4: SMS DELIVER format .....	13
Figure 2.5: First Octet for SMS SUBMIT and SMS DELIVER .....	14
Figure 2.6: TP-DCS field.....	15
Figure 2.7: STCS Format.....	15
Figure 2.8: TP-UDL and TP-UD format .....	16
Figure 3.1 : Block Diagram – Overview of Project .....	17
Figure 3.2 : The Main window .....	18
Figure 3.3 : Before generating a frequency table .....	19
Figure 3.4 : List of ASCII character code set displayed (Before generating frequency table).....	20
Figure 3.5 : Choosing a list of sample messages to generate frequency table.....	20
Figure 3.6 : Manually add a user-defined alphabet in the list.....	21
Figure 3.7 : Generating and saving the frequency table.....	21
Figure 3.8 : After generating the frequency table .....	22
Figure 3.9 : Flowchart -To generate a frequency table using the effective compression method .....	22
Figure 3.10: Block Diagram – Example to illustrate the effective compression method by reducing alphabet set size .....	23
Figure 3.11: Flowchart – To reduce the alphabet set for optimized compression.....	24
Figure 3.12: Flowchart – To generate the Huffman tree .....	25
Figure 3.13: Flowchart – To encode the alphabets .....	25
Figure 3.14: The “Send SMS” window – Part I: Compression .....	26
Figure 3.15: The arbitrary SMS message typed by user .....	27
Figure 3.16: The compressed message (in bits) .....	27
Figure 3.17: Statistics window for Compression .....	27
Figure 3.18: Hyper Terminal - Send SMS in Text Mode .....	28
Figure 3.19: Hyper Terminal - Send Single SMS in PDU Mode .....	29

Figure 3.20: An example of a Single SMS binary message .....	29
Figure 3.21: Hyper Terminal - Send Concatenated SMS in PDU Mode.....	31
Figure 3.22: Flowchart – Sending Concatenated and Single SMS through Visual Basic .....	34
Figure 3.23: The “Send SMS” window – Part II: Communication .....	35
Figure 3.24: Mobile number.....	35
Figure 3.25: PDU textbox .....	35
Figure 3. 26: Statistics for amount of SMS .....	35
Figure 3.27: Receiving Single SMS – PDU mode .....	38
Figure 3.28: Receiving Concatenated SMS – PDU mode .....	39
Figure 4.1 : Different example scenarios that will be introduced .....	40
Figure 4.2 : Relationship between Compression Ratio and Amount of reduced SMS	41
Figure 4.3 : Single SMS – Scenario A and B.....	43
Figure 4.4 : Relationship between different types of example SMS, CR and $CR_{SMS}$ ...	43
Figure 4.5 : Good SMS Example – Scenario A and B .....	45
Figure 4.6 : Average SMS Example – Scenario A and B .....	47
Figure 4.7 : Bad SMS Example – Scenario A and B .....	49
Figure 4.8 : Example used in Single SMS.....	50
Figure 4.9 : Sending Status window for Single SMS example .....	50
Figure 4.10: Received Single SMS in Hyper Terminal .....	50
Figure 4.11: Successful example of a concatenated SMS.....	51
Figure 4.12: Sending Status window for successful example of concatenated SMS .	51
Figure 4.13: Received concatenated SMS in Hyper Terminal .....	52

## List of Abbreviations

Abbreviations	Description
M2M	Machine-To-Machine
SMS	Short Message Service
SMSC	Short Message Service Centre
GSM	Global System for Mobile Communications
ETSI	European Telecommunications Standards Institute
MT	Mobile Terminal
SC	Service Centre
PC	Personal Computer
VB	Visual Basic
PDU	Protocol Data Unit
SCA	Service Centre Address
TPDU	Transport Protocol Data Unit
MR	Message Reference
DA	Destination Address
OA	Originator Address
PID	Protocol Identifier Address
DCS	Data Coding Scheme
STCS	Service Centre Time Stamp
VP	Validity Period
UDL	User Data Length
UDH	User Data Header
UD	User Data
GUI	Graphical User Interface

# Chapter 1

## Introduction

### 1.0 Overview

As the growth of mobile market increase tremendously in the recent years, Short Message Service (SMS) technology has been prevailing in many countries, including Malaysia. According to a study conducted by Malaysian Communications and Multimedia Commission (MCMC), SMS still plays a dominant role among Malaysians [1]. In particular, SMS usage reports a significant 84.9% of users in the subscriber base.

On the international level, the role of SMS as a business tool is increasing rapidly as well. Report from Cellular News in September 2006 shows that the use of SMS for business purposes was employed extensively within the South African market [2]: *"95% of respondents reported receiving business communications via SMS while 75% received email communications for business purposes. The study was conducted by Webchek using their SMS channel to conduct the research."*

However, one major drawback of using SMS is that the maximum size of one SMS is 160 characters or 140 bytes. Seeing that SMS is growing to be a common tool for business communication, there are high chances that the maximum size of one SMS will be exceeded, thus allowing more than one SMS to be combined and sent.

As such, this project caters for the needs of business enterprises to exploit SMS as an effective communication tool with other businesses and clients. In collaboration with an industrial partner, Mobitek Sdn. Bhd., The main objective is to reduce the number of SMS required to be sent by applying effective compression methods and specific rules. In particular, this project seeks to achieve a compression ratio of 50% to 60% under optimized conditions.

This project consists of two main parts:

- (i) Compression - Devise an effective compression scheme to optimize the text compression of an arbitrary SMS.
- (ii) Communication - Construct and send the appropriate format of SMS to cater for both single SMS and concatenated SMS

### 1.1 Why SMS?

Despite the long history of SMS, it is still used prevalently on the international level. The various applications of SMS is now beyond personal text messaging. It is emerging as a necessity for enterprises, where reliable and effective business to business communication tool is in great need. Currently, the main focus of this project is on the Machine-To-Machine (M2M) communications.

M2M communication makes use of mobile phone networks to exchange data two non-physically connected remote electronic devices through the SMS protocol [3]. Text message entails a piece of information using a well-defined encoding, to be properly interpreted by the receiving device. Message exchange is accomplished through SMS protocol in two ways: (a) *"Push"* technology, where device is

programmed to send SMS every time an alert is needed, or at particular time intervals, and (b) "Pull" technology, where a message reply containing desired piece of information is obtained upon sending a request message.

The main advantage of employing SMS technology is that mobile signal coverage is equal, if not superior compared to landline networks. This allows communication to devices placed in areas hardly reachable through Internet. On the other hand, it will also come in handy for devices reachable through Internet due to the following reasons:

- (i) **Reliable:** It is on very rare occasions where the mobile network is down, compared to Internet that has a higher probability of having downtime.
- (ii) **Fast:** SMS is able to reach the device/user immediately in case of any emergency.
- (iii) **Cost minimization:** Currently, the cheapest cost for SMS is as low as RM0.01

The following are some practical applications of a M2M communication:

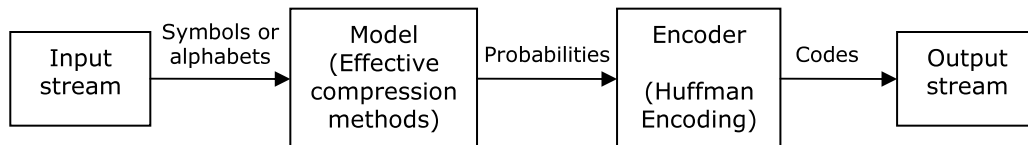
- (i) **Vending Machine management:** an automated vending machine can send every night a message containing a list of slot numbers which enclosed product was sold. Thus, it will never remain with an empty machine without products to be sold.
- (ii) **University email outage:** IT support will be able to notify all staff immediately should there be a sudden internet disruption or email outage.
- (iii) **Lift malfunction:** An alert device will immediately send a SMS to the technical assistance central.
- (iv) **ATM machine:** In the case of any problems with the ATM machines, central monitoring system will be notified immediately via SMS.
- (v) **Hospital internal communication:** In the case where urgent information of a patient is required from a specific department, the central administration will be able to receive notification and take immediate actions

## 1.2 Data Compression

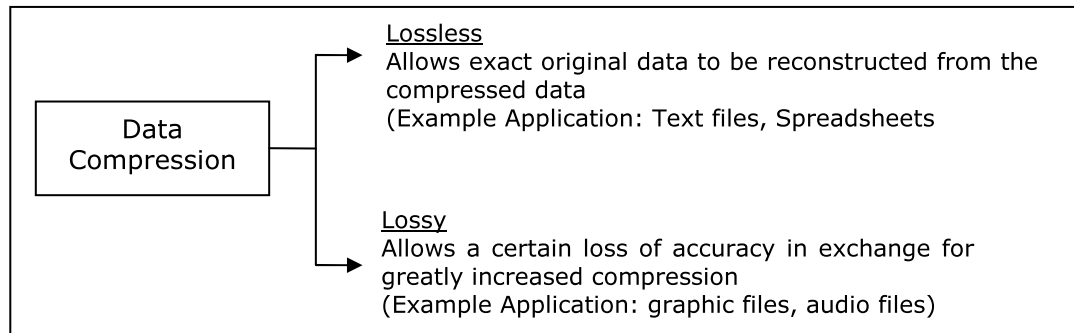
Data compression is the process of converting an input data stream into another data stream that has a smaller size [4,5]. It reduces redundancy in a message to decrease the size of the message [6]. The methods used are based on the same principle, namely, they compress data by removing redundancy from the original data. The idea behind it is to "assign short codes to common events and long codes to rare events".

$$\text{"Data Compression = Modeling + Coding"}$$

The above equation explains that data compression process consists of modeling and coding. Firstly, the decision to generate a frequency table and output a certain code for a certain symbol is based on a model. The model is simply a collection of data and rules applied on the input stream. In this project, the "model" is known as the effective compression methods, which were created to provide optimum compression. Subsequently, the input stream will undergo coding, such as Huffman coding, after being modeled. The whole data compression process is illustrated in Figure 1.1.



**Figure 1.1: The whole Data Compression process = Modeling + Coding**



**Figure 1.2: Two major families of Data Compression**

As shown in Figure 1.2, data compression techniques can be divided into two types; lossy and lossless [5]. Lossless compression generates an exact duplicate of the input data stream after compression. Thus, they are commonly used for text compression which cannot accommodate any loss of data. Hence, lossless data compression will be employed in this project since text compression was required. One of the oldest and yet commonly used lossless compression technique that will be employed in this project is Huffman coding. It is based on the statistical properties of data, i.e. probability of occurrence of each symbol in data

### 1.2.1 Huffman Coding

Huffman coding is a simple algorithm for creating variable-length codes that are an integral number of bits [5, 6]. Generally, its concept consists of the following:

- (i) Symbol that has a very high probability of occurrence generates a code with very few bits
- (ii) Symbol with a low probability of occurrence generates a code with a larger number of bits

In addition, Huffman coding, also known as an optimal compression method, has two distinct features:

- (i) **Unique Prefix:** each code is not a prefix for the subsequent one.
- (ii) **Uniquely Decodable** codes: unique input symbol for each code can be determined.

Compression ratios obtained by using Huffman coding may vary depending on how a program obtains its frequency table of the characters in the text file. Thus, in this project, it is vital to note that optimality of the compression ratios of the text messages may vary depending on how close the predicted frequency table is matches the actually frequency table from the SMS. Upon completing data compression process, the compressed message is now ready to be sent as a SMS.

One reason why Huffman coding is applied in this project is its simplicity and yet providing optimal compression. Perhaps, the most common comparison made with

Huffman is Arithmetic coding, which gives a slightly more effective compression compared to Huffman in terms of its code length [7]. However, Huffman coding is still used widespread as Arithmetic coding provides high computational complexity and patent royalties.

### 1.3 SMS

The Short Message Service (SMS) is a basic service allowing the exchange of short text messages between subscribers [8]. Initially introduced in Europe in 1992, it was included in the Global System for Mobile Communications (GSM) standards. GSM is the worldwide most popular standard which provides voice and limited data services. The format of SMS is specified by the ETSI (European Telecommunications Standards Institute) in the document GSM 03.40 [9] and GSM 03.38 [10].

The Short Message Service (SMS), as defined within the GSM digital mobile phone standard that is popular in Europe, the Middle East, Asia, Africa and some parts of North America, has several unique features [11]:

- (i) A single short message can be up to 160 characters (for a 7-bit ASCII character) of text in length. Otherwise, the maximum size will be 140 bytes for an 8-bit data set.
- (ii) The Short Message Service is a store and forward service, i.e. short messages are not sent directly from sender to recipient, but via SMS Centre (SMSC) instead. With this, SMS is able to have a special feature of having delivery status report. Unlike paging, SMS will be able to receive a confirmation message stating whether the SMS has been delivered or otherwise.
- (iii) Short messages can be sent and received simultaneously with data, GSM voice and fax calls. Rarely, users of SMS get a busy signal during peak network usage times. This is because unlike voice, data and fax calls where a dedicated radio channel was taken over, short messages travel over and above the radio channel using the signaling path.

Generally, SMS can be sent through three modes: Block mode, Text mode and PDU mode [12]. The block mode is a binary communication protocol that takes into account error protection. Text mode is a protocol that is solely based on AT commands and PDU mode is a protocol that is based on hexadecimal encoded binary data that contains command and useful information.

In this project, compressed message, in the form of binary data, was sent as a SMS. As such, the PDU mode was chosen, with the maximum size of a SMS of 140 bytes (8-bit data). In the case where the SMS size exceeds 140 bytes, SMS concatenation will occur, i.e. linking several messages together.

## 1.4 Sending SMS through GSM modem

Upon constructing the PDU format from the SMS, it is now ready to be sent through GSM modem. Figure 1.3 shows the GSM modem that will be used throughout the project. Prior to communicating with the GSM modem, AT commands were used to communicate with the GSM modem. AT commands are a set of commands that are used to for communication between Personal Computer (PC) to modem, and vice versa. There is a Communications program in Microsoft® Windows that allows user to verify the AT commands used before implementing it in applications and project.



Figure 1.3: Wavecom Fastrack modem M1206B

## 1.5 Project Outline

Following the introduction to the essentials of information required, this section describes the outline of the project, which further elaborates the overview in 1.0.

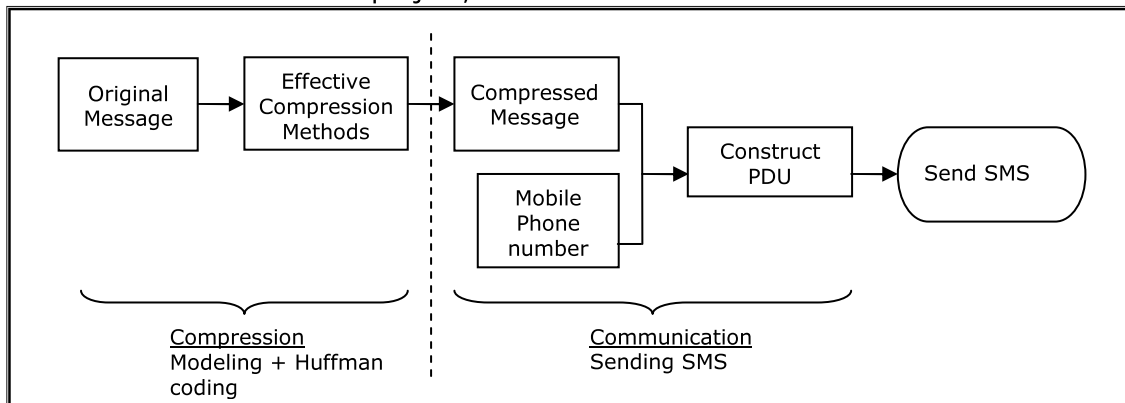


Figure 1.4: Block Diagram - Project Outline

The following describes the block diagram in Figure 1.4:

- 1) Apply the effective compression methods to the original SMS to shape the predicted frequency table close to the actual frequency table. This aids in producing an optimum compression ratio.
- 2) Upon preparing the best possible frequency table (probability set), original message is replaced with the Huffman codes to produce compressed message in the form of binary data.
- 3) Compressed message, along with the mobile phone number, are converted into the appropriate PDU format.
- 4) Connection between GSM modem and PC is established through the usage of AT commands.
- 5) SMS is now ready to be sent either in one SMS or concatenated SMS.

## **1.6 Proposed Deliverables**

Upon completion of the project and achieving the main objective described in 1.0, the following proposed deliverables are to be met:

- (a) Creating user-friendly Visual Basic (VB) interfaces for:
  - i. User – Type an arbitrary SMS and send it
  - ii. Administrator – To collect and process sample messages for frequency table generation
- (b) Able to send SMS via a modem in PDU format through the VB interface
- (c) Write long arbitrary text messages and achieve effective compression that can be fitted into fewer text messages.

## **1.7 Thesis Structure**

This chapter has introduced the motivation underlying the work presented, objectives and proposed outcomes upon completing the project. In addition, a brief background information and characteristics of the required knowledge were introduced. The rest of this thesis will be devoted to the development of an effective text compression through Visual Basic interface and the evaluation of the results obtained.

Chapter 2 provides the necessary knowledge and information required to understand and complete the project.

Chapter 3 elaborates the process of implementing the effective text compression method and subsequently sending the SMS. The whole process was implemented through Visual Basic interface.

Chapter 4 seeks to display the results obtained from effective compression, and evaluate the optimality of the results. It further verifies that all the SMS sent were able to be received accurately.

Finally, Chapter 5 provides a conclusion of the thesis and the possible future development of the project.