SMS character set handling and multipart messages

This guide gives you information about message length and character set handling of SMS messages. This is an important topic concerning costs. Please read through it to make sure you are able to configure the system according to your preferences.

Introduction

SMS is used to send text messages between mobile phones in most cases. When a text is transmitted, there is a size limitation on the message length. If English characters are used, the maximum message length is 160 characters. When international characters are sent, the maximum length is 70 characters. This size limit is determined by the character set used to transmit the message.

SMS segmentation and reassembly (SAR)

To increase the size limit of text messages, the SMS technology was improved to support longer text messages. This improvement is called as the multipart SMS technology, which refers to a process known as segmentation and reassembly procedure.
If an English text message is longer than 160 characters then, it is first segmented by the sending mobile and is then transmitted through the GSM network in several SMS messages. The recipient mobile phone, after receiving all message parts reassembles the segments and displays the long text as a single message to the user.

If international characters are used the segmentation starts when a message text becomes longer than 70 characters.

When multipart technology is applied, the cost of each message can be calculated by the number SMS messages used to transmit the text over the wireless network. For example if a text message is 240 English characters it fits into two SMS, so the cost will be twice as much as a single 160 character SMS.

Reduced Message length for Messages over 160 characters:

One might expect that if a single SMS can hold 160 characters, a 320 character message would take two physical SMS Messages. This is not the case, when multipart SMS technology is used only 153 characters fit into a single SMS, because some space is needed for the segmentation information that is used to reassemble the message parts in correct order. So if a 320 character message is sent, it would take 3 SMS. The first two would hold 153 characters, and the last one would hold 14 characters. For international characters, 67 characters fit into a multipart SMS segment.

Terms and definitions, SAR technology in detail

Character Sets and Unicode:

English Character Set:

The English characters set refers to the 7 bit SMS alphabet, that contains English characters and a few international characters for Western Europe and Greece. These characters are defined in the ETSI GSM 03.38 standard.

International Character Set:

International characters or Unicode character set. The Unicode character set can be used to send special symbols and characters of all languages including Chinese, Arabic, Hebrew, Cyrillic, special eastern European characters, etc.

In GSM SMS system, an SMS message can contain up to 140 bytes (standard 8-bit bytes) of message data. The 7 bit SMS alphabet makes it possible to send 160 characters in these 140 bytes.
This means that, when you send a text message, as long as the text only contains characters that are included in the GSM 7 bit SMS alphabet character set, 160 7-bit characters are compressed into 140 8-bit bytes to produce the 160 character limit that we are so familiar with. (Note: 160 * 7 = 140 * 8).

It is worth noting that ETSI GSM 03.38 also defines a few characters that are represented by two 7-bit characters when included in a text message.
Ref: Appendix A: "^", "{", "}", "\", "[", "]", "~", "" and "€".

If you want to send a message that contains characters that are not part of the GSM 7-bit character set, such as Chinese, Arabic, Thai, Cyrillic, etc., then the entire text of the SMS that actually goes out over the air needs to be encoded in the Unicode UCS-2 character set.
In the UCS-2 character set, each character is encoded with 16-bits (or two 8-bit bytes). This means that an SMS message is limited to 70 16-bit Unicode characters (70 * 16 = 140 * 8).

If a message is larger than 140 8-bit bytes, then segmentation and reassembly standards define, that where a single logical message can be sent over the air using multiple physical SMS messages. The receiving client then has the ability to reassemble the segmented message so that it again appears as a single message on the receiving device.

When a long text message is segmented into multiple physical SMS messages, a special header is added to each physical SMS message so that the receiving client knows that it is a multipart SMS message that must be reassembled by the client. These headers are known as segmentation or concatenation headers or SAR headers. The SAR headers are 6 bytes (8-bits each). They are included in each physical SMS message. These headers are placed in the User Data Header (UDH) field of the message, but they do count against the overall size limit of the message.

If you send a long text message containing only characters that are part of the GSM 03.38 character set, then each SMS segment can contain up to 153 characters. (140 bytes - 6 bytes for the concatenation header leaves 134 available bytes, or 7 * 134 = 1072 bits. The most 7-bit characters that can be packed into 1072 bits is 153.)

If you send a long text message that includes any characters that require Unicode encoding, then each SMS segment can contain up to 67 characters. (67 * 16 = 1072 bits)

Appendix A - The 7 bit default alphabet of GSM phones

This table gives you information about the GSM 7 bit alphabet used in text SMS messages. To see how you can send special symbols and international characters, please read the following document:

Your PC and your GSM phone use two different character sets: the ISO-8859-1 and the GSM 7 bit alphabet.

This is the 7 bit default alphabet as specified by GSM 03.38. The corresponding ISO-8859-1 decimal codes are shown in the rightmost column. Note that the euro sign (€) is included in the ISO-8859-15 character set.

Hex	Dec	Character name	Character	ISO-8859-1 DEC
0x00	0	COMMERCIAL AT	@	64
0x01	1	POUND SIGN	£	163
0x02	2	DOLLAR SIGN	$	36
0x03	3	YEN SIGN	¥	165
0x04	4	LATIN SMALL LETTER E WITH GRAVE	è	232
0x05	5	LATIN SMALL LETTER E WITH ACUTE	é	233
0x06	6	LATIN SMALL LETTER U WITH GRAVE	ù	249
0x07	7	LATIN SMALL LETTER I WITH GRAVE	ì	236
0x08	8	LATIN SMALL LETTER O WITH GRAVE	ò	242
0x09	9	LATIN CAPITAL LETTER C WITH CEDILLA	Ç	199
0x0A	10	LINE FEED		10
0x0B	11	LATIN CAPITAL LETTER O WITH STROKE	Ø	216
0x0C	12	LATIN SMALL LETTER O WITH STROKE	ø	248
0x0D	13	CARRIAGE RETURN		13
0x0E	14	LATIN CAPITAL LETTER A WITH RING ABOVE	Å	197
0x0F	15	LATIN SMALL LETTER A WITH RING ABOVE	å	229
0x10	16	GREEK CAPITAL LETTER DELTA	Δ	
0x11	17	LOW LINE	_	95
0x12	18	GREEK CAPITAL LETTER PHI	Φ	
0x13	19	GREEK CAPITAL LETTER GAMMA	Γ	
0x14	20	GREEK CAPITAL LETTER LAMBDA	Λ	
0x15	21	GREEK CAPITAL LETTER OMEGA	Ω	
0x16	22	GREEK CAPITAL LETTER PI	Π	
0x17	23	GREEK CAPITAL LETTER PSI	Ψ	
0x18	24	GREEK CAPITAL LETTER SIGMA	Σ	
0x19	25	GREEK CAPITAL LETTER THETA	Θ	
0x1A	26	GREEK CAPITAL LETTER XI	Ξ	
0x1B	27	ESCAPE TO EXTENSION TABLE		
0x1B0A	27 10	FORM FEED		12
0x1B14	27 20	CIRCUMFLEX ACCENT	^	94
0x1B28	27 40	LEFT CURLY BRACKET	{	123
0x1B29	27 41	RIGHT CURLY BRACKET	}	125

0x1B2F	27 47	REVERSE SOLIDUS (BACKSLASH)	\	92
0x1B3C	27 60	LEFT SQUARE BRACKET	[	91
0x1B3D	27 61	TILDE	~	126
0x1B3E	27 62	RIGHT SQUARE BRACKET	]	93
0x1B40	27 64	VERTICAL BAR	|	124
0x1B65	27 101	EURO SIGN	€	164 *(ISO-8859-15)*
0x1C	28	LATIN CAPITAL LETTER AE	Æ	198
0x1D	29	LATIN SMALL LETTER AE	æ	230
0x1E	30	LATIN SMALL LETTER SHARP S (German)	ß	223
0x1F	31	LATIN CAPITAL LETTER E WITH ACUTE	É	201
0x20	32	SPACE		32
0x21	33	EXCLAMATION MARK	!	33
0x22	34	QUOTATION MARK	"	34
0x23	35	NUMBER SIGN	#	35
0x24	36	CURRENCY SIGN	¤	164 (ISO-8859-1)
0x25	37	PERCENT SIGN	%	37
0x26	38	AMPERSAND	&	38
0x27	39	APOSTROPHE	'	39
0x28	40	LEFT PARENTHESIS	(	40
0x29	41	RIGHT PARENTHESIS	)	41
0x2A	42	ASTERISK	*	42
0x2B	43	PLUS SIGN	+	43
0x2C	44	COMMA	,	44
0x2D	45	HYPHEN-MINUS	-	45
0x2E	46	FULL STOP	.	46
0x2F	47	SOLIDUS (SLASH)	/	47
0x30	48	DIGIT ZERO	0	48
0x31	49	DIGIT ONE	1	49
0x32	50	DIGIT TWO	2	50
0x33	51	DIGIT THREE	3	51
0x34	52	DIGIT FOUR	4	52
0x35	53	DIGIT FIVE	5	53
0x36	54	DIGIT SIX	6	54
0x37	55	DIGIT SEVEN	7	55
0x38	56	DIGIT EIGHT	8	56
0x39	57	DIGIT NINE	9	57
0x3A	58	COLON	:	58
0x3B	59	SEMICOLON	;	59
0x3C	60	LESS-THAN SIGN	<	60
0x3D	61	EQUALS SIGN	=	61
0x3E	62	GREATER-THAN SIGN	>	62
0x3F	63	QUESTION MARK	?	63
0x40	64	INVERTED EXCLAMATION MARK	¡	161
0x41	65	LATIN CAPITAL LETTER A	A	65
0x42	66	LATIN CAPITAL LETTER B	B	66

0x43	67	LATIN CAPITAL LETTER C	C	67
0x44	68	LATIN CAPITAL LETTER D	D	68
0x45	69	LATIN CAPITAL LETTER E	E	69
0x46	70	LATIN CAPITAL LETTER F	F	70
0x47	71	LATIN CAPITAL LETTER G	G	71
0x48	72	LATIN CAPITAL LETTER H	H	72
0x49	73	LATIN CAPITAL LETTER I	I	73
0x4A	74	LATIN CAPITAL LETTER J	J	74
0x4B	75	LATIN CAPITAL LETTER K	K	75
0x4C	76	LATIN CAPITAL LETTER L	L	76
0x4D	77	LATIN CAPITAL LETTER M	M	77
0x4E	78	LATIN CAPITAL LETTER N	N	78
0x4F	79	LATIN CAPITAL LETTER O	O	79
0x50	80	LATIN CAPITAL LETTER P	P	80
0x51	81	LATIN CAPITAL LETTER Q	Q	81
0x52	82	LATIN CAPITAL LETTER R	R	82
0x53	83	LATIN CAPITAL LETTER S	S	83
0x54	84	LATIN CAPITAL LETTER T	T	84
0x55	85	LATIN CAPITAL LETTER U	U	85
0x56	86	LATIN CAPITAL LETTER V	V	86
0x57	87	LATIN CAPITAL LETTER W	W	87
0x58	88	LATIN CAPITAL LETTER X	X	88
0x59	89	LATIN CAPITAL LETTER Y	Y	89
0x5A	90	LATIN CAPITAL LETTER Z	Z	90
0x5B	91	LATIN CAPITAL LETTER A WITH DIAERESIS	Ä	196
0x5C	92	LATIN CAPITAL LETTER O WITH DIAERESIS	Ö	214
0x5D	93	LATIN CAPITAL LETTER N WITH TILDE	Ñ	209
0x5E	94	LATIN CAPITAL LETTER U WITH DIAERESIS	Ü	220
0x5F	95	SECTION SIGN	§	167
0x60	96	INVERTED QUESTION MARK	¿	191
0x61	97	LATIN SMALL LETTER A	a	97
0x62	98	LATIN SMALL LETTER B	b	98
0x63	99	LATIN SMALL LETTER C	c	99
0x64	100	LATIN SMALL LETTER D	d	100
0x65	101	LATIN SMALL LETTER E	e	101
0x66	102	LATIN SMALL LETTER F	f	102
0x67	103	LATIN SMALL LETTER G	g	103
0x68	104	LATIN SMALL LETTER H	h	104
0x69	105	LATIN SMALL LETTER I	i	105
0x6A	106	LATIN SMALL LETTER J	j	106
0x6B	107	LATIN SMALL LETTER K	k	107
0x6C	108	LATIN SMALL LETTER L	l	108
0x6D	109	LATIN SMALL LETTER M	m	109
0x6E	110	LATIN SMALL LETTER N	n	110
0x6F	111	LATIN SMALL LETTER O	o	111

0x70	112	LATIN SMALL LETTER P	p	112
0x71	113	LATIN SMALL LETTER Q	q	113
0x72	114	LATIN SMALL LETTER R	r	114
0x73	115	LATIN SMALL LETTER S	s	115
0x74	116	LATIN SMALL LETTER T	t	116
0x75	117	LATIN SMALL LETTER U	u	117
0x76	118	LATIN SMALL LETTER V	v	118
0x77	119	LATIN SMALL LETTER W	w	119
0x78	120	LATIN SMALL LETTER X	x	120
0x79	121	LATIN SMALL LETTER Y	y	121
0x7A	122	LATIN SMALL LETTER Z	z	122
0x7B	123	LATIN SMALL LETTER A WITH DIAERESIS	ä	228
0x7C	124	LATIN SMALL LETTER O WITH DIAERESIS	ö	246
0x7D	125	LATIN SMALL LETTER N WITH TILDE	ñ	241
0x7E	126	LATIN SMALL LETTER U WITH DIAERESIS	ü	252
0x7F	127	LATIN SMALL LETTER A WITH GRAVE	à	224