데이터통신 전송효율과 라틴어 부호 체계 고찰

홍완표*

A study on Code System of Latin Character to Improve Transmission Efficiency in Data Communications

Wan-Pyo Hong*

요 약

본 논문은 국제표준 문자부호 체계인 유니코드(Unicode) 3.0에 포함된 라틴어 문자에 관한 새로운 부호 체계를 제시하였다. 이 연구의 시작 배경은 Unicode 3.0의 라틴어 부호 체계가 데이터통신의 전송효율 측면에서 적정한가에 대한 것이었다. 데이터통신을 할 때, 4개 또는 8개 이상의 연속 "0"의 비트가 단말 정보기기로 부터 회선부호화 장치에 입력 될 수 있다. 이 경우에 그 비트열은 스크램블링 과정을 거쳐 연속 "0"이 아닌 비트열로 변경된다. 그러므로 단말 정보기기에서 처리되는 0 문자, 기호 등의 부호 체계에 따라서 데이터통신의 회선부호기 운용이 달라지게 된다. 즉, 데이터통신의 전송효율에 영향을 주게 된다. 이러한 관점에서 본 논문은 [1]에서 제시된 (4 x 4)hexa 원천 부호화 규칙과 영어 문자의 사용빈도 통계를 적용하여 유니코드와 UTF-8의 라틴어 부호 체계에 대한 개선방안을 제시하였다. 그 결과 본 연구에서 제시한 개선된 유니코드와 UTF-8 라틴어 부호 체계를 적용할 경우, 회선부호기의 스크램블러 운용효율이 유니코드를 통신용으로 사용할 경우 최소 3645%에서 최대 31400%, 제시된 UTF-8 부호 체계를 적용할 경우 최소 480%에서 최대 1700%까지 개선되는 것으로 나타났다.

ABSTRACT

This paper proposes the revised Roman character code system using Unicode 3.0. The background of the paper is whether the Latin character code system th using in the world in Unicode V.3 is proper or not in the side of the transmission efficiency in data communications. In data communications, when the consecutive 4 bits or 8 bits of "0" bit from the information devices input into the line coder, its consecutive "0" bits are scrambled to the predetermined bit patterns to avoid the syncronization loss. The paper was based on the statistical data for the using frequency of the alphabet letter and the proposed rule for characters coding in [1]. The paper was focused to improve of Unicode itself and UTF-8 code system. As a result of the paper, when the proposed coding systems for Latin character in Unicode 3.0 itself and UTF-8 code system, the scrambler efficiency using HDB-3 in the line coder of the data transmission system could be improved about 3645 ~ 31400% and 480~1700% respectively.

키워드

Unicode, Line coder, HDB-3, Scrambling
유니코드, 회선부호기, HDB-3, 스크램블링

* 한세대학교 정보통신공학과(wphong@hansei.ac.kr)

접수일자 : 2012. 06. 13 심사(수정)일자 : 2012. 07. 26 게재확정일자 : 2012. 08. 09.

Ⅰ. 서 론

본 논문에서는 컴퓨터 등 정보 단말 기기의 원천부호화(source coding) 체계에 대하여 연구하였다. 정보 단말 기기의 입력기로부터 입력되는 문자, 기호 등 정보는 정해진 부호로 바뀌어 저장되거나 출력장치로 출력된다. 또한 이 부호화된 정보는 모뎀, LAN 카드 등의 통신장치를 이용하여 통신망을 통해 원거리에 전송된다. 이와 같이 정보 단말 기기내에서 문자, 기호등의 부호 체계는 ASCII[2][3][4], EBCDIC[5], Unicode[6] 등이 있다. 이 부호 체계 중에서 ASCII와 EBCDIC부호 체계는 한 개의 문자나 기호가 각각 7비트, 8비트로 부호화된다. 이러한 부호 체계로는 전 세계의 다양한 문자와 기호를 수용할 수 없다. 기본 유니코드는 한 개의 문자나 기호를 32비트로 부호화함으로서 이들 문제를 해결하고자 하는 부호 체계이다. 문자나 기호 등의 이진비트 부호는 조합형 또는 완성형의 형태로 처리된다. 라틴어는 조합형이고 한글은 조합형[7]과 완성형[8] 코드로 되어 있다. 본 논문에서는 AMI회선 부호화 방식과 이를 보강하기 위한 스크램블 방식으로 ITU-T[9] 및 국내의 표준방식[10]인 HDB-3 스크램블방식을 바탕으로 연구하였다. 회선부호화는 정보 기기 내에서 생성된 원천부호가 전송로를 통하여 원격지에 전송되기 위해 전송에 적합한 신호로 부호화하는 것이다. 연속되는 “0”의 비트 신호 열이 전송로 상에 전송될 경우 수신기의 역부호기(decoder)에서 비트 간 동기를 정확히 맞출 수 없게 된다. 즉 4개 이상의 연속된 “0”비트의 신호 열은 수신기의 복호기에서 정확한 데이터를 검파하지 못하게 할 수 있다. HDB-3 스크램블 방식은 4개의 연속된 이진비트 “0”이 회선 부호기에 입력되면 이 4개의 이진비트 “0”을 “0”레벨의 연속이 아닌 사전에 정해진 신호패턴으로 변경시킨다[11]. 결과적으로 컴퓨터 등 정보기기에서 문자나 기호 등을 어떠한 규칙에 따라 부호화하는 가에 따라서 전송 회선부호기와 복호기에서의 데이터 처리의 효율성에 영향을 주게 된다. 본 논문은 [1]에서 기 제시된 4 비트 x 4비트 원천 부호화 규칙을 적용한 수정된 기본 유니코드 라틴어 부호 체계를 제시하였다. 또한 본 논문에서 제시한 수정된 기본 유니코드 라틴어 부호 체계를 적용할 경우 예상되는 회선부호기에서의 데이터처리 효율의 향상 정도를 제시하였다.

Ⅱ. 유니코드(Unicode) 부호 체계

2.1 개요

초기의 유니코드는 문자나 기호가 16비트(2바이트 x 8비트) 체계였다. 그러므로 이 부호 체계는 총65,536개의 문자, 기호를 수용할 수 있다. 이 부호 체계로는 전 세계의 다양한 언어와 기호를 수용할 수 없게 되었다. 현재의 유니코드는 표 1과 같은 유니코드 버전3.0 체계이다. 이 유니코드 버전3.0은 32비트(4바이트 x 8비트) 체계를 갖는다. 이 4바이트 체계는 상위 2바이트와 하위2바이트로 구성된다. 상위 2바이트는 유니코드 문자판(plane)을 나타낸다. 그러므로 총 문자판은 65,536개가 된다. 하위2바이트는 문자와 기호에 부여하는 것으로 각 문자판마다 총 65,536개의 문자나 기호에 대한 부호를 부여할 수 있다. 결과적으로 유니코드 버전3은 총 4,294,967,296개의 문자와 기호를 수용할 수 있는 부호 체계이다[12]. 표 1에서 BMP(Basic Multilingual Plane)는 기본 다국어 문자판을 나타내는 것으로 상위 2바이트의 구성이 “0000”이 된다. 즉 상위 2바이트 16비트는 모두 “0”의 값을 갖는다. 이 기본다국어 문자판은 기존의 2바이트 유니코드를 수용하고 있다.

SMP(Supplementary Multilingual Plane)는 추가 다국어 문자판으로 BMP에 포함되지 않은 다국어 문자에 대한 추가코드들을 제공한다. SIP(Supplementary Ideographic Plane)는 한자 등 상형문자, 기호 그리고 발음과 대조를 이루는 의미를 표시하기 위한 기호에 대한 코드를 제공한다. SSP(Supplementary Special Plane)는 특수문자에 대한 코드를 제공한다. PUA(Private Use Area)는 사용자가 개인적으로 코드를 부여하도록 제공한다. 표 1에서 정의되지 않은 0003-000D, 0011-FFFF 문자판들은 아직 코드가 부여되지 않은 여유문자판들이다[13][14].

표 1. 유니코드 3.0 체계
Table 1. Unicode 3.0 system

BMP		SMP		SIP		SSP	PUA			
0000 - 0FFF	8000 - 8FFF	10000 - 10FFF	18000-18FFF	20000 - 20FFF	28000 - 28FFF	E0000 - E0FFF	F0000 - F0FFF	F8000 - F8FFF	100000 - 100FFF	108000 - 108FFF
1000 - 1FFF	9000 - 9FFF	11000 - 11FFF	19000-19FFF	21000 - 21FFF	29000 - 29FFF		F1000 - F1FFF	F9000 - F9FFF	101000 - 101FFF	109000 - 109FFF
2000 - 2FFF	A000 - AFFF	12000 - 12FFF	1A000-1AFFF	22000 - 22FFF	2A000 - 2AFFF		F2000 - F2FFF	FA000 - FAFFF	102000 - 102FFF	10A000 - 10AFFF
3000 - 3FFF	B000 - BFFF	13000 - 13FFF	1B000-1BFFF	23000 - 23FFF	2B000 - 2BFFF		F3000 - F3FFF	FB000 - FBFFF	103000 - 103FFF	10B000 - 10BFFF
4000 - 4FFF	C000 - CFFF	14000-14FFF	1C000-1CFFF	24000 - 24FFF	2C000 - 2CFFF	F4000 - F4FFF	FC000 - FCFFF	104000 - 104FFF	10C000 - 10CFFF	
5000 - 5FFF	D000 - DFFF	15000-15FFF	1D000 - 1DFFF	25000 - 25FFF	2D000 - 2DFFF	F5000 - F5FFF	FD000 - FDFFF	105000 - 105FFF	10D000 - 10DFFF	
6000 - 6FFF	E000 - EFFF	16000 - 16FFF	1E000-EFFF	26000 - 26FFF	2E000 - 2EFFF	F6000 - F6FFF	FE000 - FEFFF	106000 - 106FFF	10E000 - 10EFFF	
7000 - 7FFF	F000 - FFFF	17000-17FFF	1F000 - 1FFFF	27000 - 27FFF	2F000 - 2FFFF		F7000 - F7FFF	FF000 - FFFFF	107000 - 107FFF	10F000 - 10FFFF

표 2. Unicode version과 문자수(개)
Table. 2 Unicode version verse number of character

Version(일자)	문자(개)
1.0 (1991.10)	7,161
1.1(1993. 7)	34,233
2.0(1996. 7)	38,950
3.0(1999. 9)	49,259
3.1(2001. 3)	94,205
3.2(2002. 3)	95,221
4.0(2003. 4)	96,447
4.1(2005. 3)	97,720
5.0(2006. 7)	99,089
5.1(2008. 3)	100,713
5.2(2009.10)	107,361
6.0(2010.10)	109,449
6.1(2012.1)	110,181

표 2는 유니코드의 버전과 각 버전에서 수용하고 있는 문자의 수를 보여주고 있다[15]. 표 2에서 보는 것과 같이 1991년 버전1.0에 부여된 문자수와 2001년 버전3.0에 부여된 문자수 간에는 무려 13배의 증가가 있었음을 알 수 있다. 반면에 2001년 버전3.0과 2012년 버전 6.1간에는 1.6배의 증가가 있었음을 알 수 있다. 현재 전 세계에는 약 6,000개의 언어가 존재하는데, 그 중 90~95%가 21세기 중에 소멸될 것으로 내다본다[16]. 이것은 언어의 공용어 등에 의한 소수언어가 소멸되기 때문으로 보고 있다. 이러한 현상은 결과적으로 장기적인 관점에서 볼 때 유니코드에 부여될 언어의 문자수가 급격히 증가할 것으로는 예상되지 않는다.

2.2 유니코드 1.0 라틴어 부호 체계

표 3은 유니코드 1.0에 있는 확장ASCII 부여체계이다. 이 표 3에서 보는 바와 같이 7비트 1바이트의 ASCII부호가 2바이트 16비트로 확장된 것을 알 수 있다. 즉 ASCII 7비트를 구성하는 상위 3비트에 한 개의 “0”비트가 추가되어 상위비트가 4비트가 되었다. 이와 같이 구성된 8비트에 8비트 1개 바이트(x, x) hexa)가 추가되어 16비트가 된 것이다. 여기서 추가된 1바이트는 전체 2바이트 중 상위바이트가 된다. 따라서 상위바이트의 범위는 00~FF가 된다. 표 3의 각 부호에는 이 범위 중에서 00이 상위바이트로 구성되어 있음을 알 수 있다. 표 1에서와 같이 유니코드 3.0은 4바이트 32비트로 구성된다. 그러므로 표 3의 2바이트에 2바이트가 추가되어야 한다. 이렇게 추가되는 2바이트는 전체 4바이트의 상위 바이트가 된다. 이 상위 2바이트는 표 1에서 설명한 바와 같이 유니코드

표 3. 유니코드 1.0 C0 제어와 기본 라틴어 부호 체계
Table 3. C0 controls and basic latin code system in unicode 1.0

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0000	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
0010	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
0020	SP	!	″	#	$	%	&	′	(	)	*	+	,	-	.	/
0030	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
0040	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
0050	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
0060	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
0070	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

3.0의 문자판을 식별하는 바이트가 된다. 표 3의 확장 ASCII 부호의 상위바이트는 표 1의 BMP범위에 속한다. 그러므로 표 3의 각 문자와 기호 등을 나타내는 모든 부호에는 0000 hexa의 2바이트가 추가된다. 예를 들어 영문자 대문자 A의 부호는 00000041이 되고 소문자 a는 00000061이 된다. 표 3에서 SOH, STX, ETX, EOT, ENQ, ACK, BEL 등은 통신프로토콜에서 사용되는 문자들이다. 즉 이 문자들은 회선부호기의 스크램블러를 거쳐야 하는 문자들이다. 반면에 BS, HT, LF, VT, FF, CR, SO, SI, DEL, DC, ESC, FS, GS, RS, US, SP 등은 정보기기 자체 내에서 제어 등에 사용되는 문자들이다.

III. 라틴어의 사용빈도에 대한 통계

표 4는 라틴어에 대한 사용빈도를 보여주는 것으로 모르스 부호, 옥스퍼드 사전 및 실용단어의 사용빈도이다. 모르스 부호의 발명가인 사무엘 모르스(Samuel Morse; 1791-1872)는 사용빈도가 높은 글자를 체계가 가장 간단한 부호에 부여하고자 하였다[17]. 이것을 위해서는 라틴어 문자의 사용 빈도수를 조사해야만 했다. 모르스는 단순한 평문으로 구성된 문서를 분석하여 상대적으로 사용빈도수가 많은 라틴어를 조사하였다. 모르스에 의한 라틴어의 사용빈도 조사는 평서문을 근간으로 하고 있다. 그러므로 실제 다양한 단어가 사용되는 문서에서의 라틴어 사용빈도수는 이와 다르게 나타난다.

옥스퍼드 사전에 의한 사용빈도는 컨사이스 옥스퍼드사전 (Concise Oxford Dictionary)의 단어 목록의 글자로부터 분석된 것이다[17].

표 4. 라틴어 사용빈도에 대한 모르스 부호, 옥스퍼드 사전, 실용500단어 및 전체평균
Table 4. Usage frequency of morse code, oxford dictionary, practical english and total average in latin

모르스 부호		옥스퍼드 사전		실용500 단어		평균	
E	17.83	E	11.16	E	11.7	E	13.57
T	13.37	A	8.50	A	7.91	A	9.43
A	11.89	R	7.58	O	7.87	T	9.25
I	11.89	I	7.54	T	7.43	O	8.97
N	11.89	O	7.16	R	7.33	N	8.33
O	11.89	T	6.95	N	6.45	I	8.25
S	11.89	N	6.65	S	5.52	R	8.04
H	9.51	S	5.74	I	5.32	S	7.71
R	9.21	L	5.49	L	5.18	H	5.72
D	6.54	C	4.54	H	4.64	L	5.54
L	5.94	U	3.63	D	4.01	D	4.64
U	5.05	D	3.38	U	3.57	C	4.09
C	4.46	P	3.17	C	3.27	U	4.08
M	4.46	M	3.01	W	2.93	M	3.45
F	3.71	H	3.00	M	2.88	P	2.69
W	2.97	G	2.47	G	2.59	F	2.61
Y	2.97	B	2.07	P	2.39	G	2.53
G	2.53	F	1.81	F	2.30	W	2.40

P	2.53	Y	1.78	Y	2.00	Y	2.25
B	2.38	W	1.29	B	1.86	B	2.10
V	1.78	K	1.10	V	1.22	V	1.34
K	1.19	V	1.01	K	1.17	K	1.15
Q	0.74	X	0.29	X	0.20	X	0.36
J	0.59	Z	0.27	J	0.10	Q	0.35
X	0.59	J	0.20	Q	0.10	J	0.30
Z	0.30	Q	0.20	Z	0.05	Z	0.21

실생활사용단어의 사용빈도는 영국, 미국 및 호주의 영자신문, 잡지, 서적, TV, 라디오 및 실생활회화에서 사용하는 영어에서 가장 많이 사용되는 500단어에서 추출한 것이다[18].

표 5. 각 사용빈도 통계치의 유사성 비교
Table 5. Comparison of statistical similarity between each usage frequency rate

(a) 사용빈도 평균 5% 이상의 경우
(a) In case of over 5% of average usage frequency

모르스 부호		옥스포드 사전		실용 500단어		평 균	
D	6.54	C	4.54	E	11.72	E	13.57
E	17.83	E	11.16	H	4.64	H	5.72
H	9.51	I	7.54	I	5.32	I	8.25
I	11.89	L	5.49	L	5.18	L	5.54
N	11.89	N	6.65	N	6.45	N	8.33
O	11.89	O	7.16	O	7.87	O	8.97
R	9.21	R	7.58	R	7.33	R	8.04
S	11.89	S	5.74	S	5.52	S	7.71
T	13.37	T	6.95	T	7.43	T	9.25

(b) 사용빈도 평균 8% 이상
(b) In case of over 8% of average usage frequency

모르스 부호		옥스포드 사전		실용 500단어		평 균	
E	17.83	E	11.16	E	11.72	E	13.57
I	11.89	I	7.54	N	6.45	I	8.25
N	11.89	N	6.65	O	7.87	N	8.33
O	11.89	O	7.16	R	7.33	O	8.97
S	11.89	R	7.58	S	5.52	R	8.04
T	13.37	T	6.95	T	7.43	T	9.25

(c) 사용빈도 평균 1% 이하
(c) In case of below 1% of average usage frequency

모르스 부호		옥스포드 사전		실용 500단어		평 균	
K	1.19	K	1.10	K	1.17	K	1.15
Q	0.74	Q	0.20	Q	0.10	Q	0.35
V	1.78	V	1.01	V	1.22	V	1.34
X	0.59	X	0.29	X	0.20	X	0.36
Z	0.30	Z	0.27	Z	0.05	Z	0.21

표 5 (a)는 평균 사용빈도의 경우에 빈도율이 5이상인 경우에 대한 것이다. 평균 사용빈도를 기준으로 하였을 때 모르스 부호와 옥스퍼드사전이 각각 한 개씩 틀림을 알 수 있다. 평균과 실용단어 500개와는 모두 동일함을 알 수 있다. 표 5(b)의 경우에는 평균 사용빈도의 빈도율이 8%이상인 경우에 대한 것이다. 평균 사용빈도를 기준으로 하였을 때 옥스퍼드 사전과는 동일하고 나머지 두 개와는 한 개씩 틀림을 알 수 있다. 표 5(c)의 경우에는 평균 사용빈도의 빈도율이 1%이하인 경우에 대한 것이다. 이 표 5에서와 같이 사용빈도가 1%이하인 경우에는 네 개의 사용빈도 모두가 동일함을 알 수 있다. 그러므로 본 논문에서는 표 4의 평균값을 적용하였다.

Ⅳ. 유니코드 1.0 라틴어부호 체계 개선방안

4.1 4x4비트 원천부호화규칙

표 6은 참고문헌[1]에서 제시한 4비트 x 4비트 원천 부호화 규칙이다. 이 표 6에서 제시하는 원천 부호화 원칙은 8비트열 중에서 상위 4비트열을 기준으로 한 것이다. 이 원칙을 2바이트 16비트열에 적용할 경우, 16비트열의 최상위 4비트를 기준으로 적용하게 된다. 이것이 4바이트 32비트열에 적용된다면 32비트열의 최상위 4비트가 표 3의 최상위 비트열이 된다. 그러므로 8바이트 32비트 유니코드가 그대로 통신신호로 전송된다고 가정할 경우, 표 6의 원천 부호화 규칙에 의하여 첫 번째로 사용을 피하여 할 최상위 4비트열은 “0000”이 된다. 즉, 유니코드 3.0의 문자판을 식별하는 2바이트 중 상위바이트 8비트중 상위4비트열이 0 hexa가 되는 문자판의 사용을 피해야 한다. 두 번째 사용을 피하여야 하는 최상위 4비트열은 8 hexa

가 된다. 세 번째로는 (4, C)hexa가 된다. 네 번째로는 (2, 6, A, E)hexa가 된다. 결과적으로 이상의 것을 제외한 (1, 3, 5, 7, 9, B, D, F)hexa 가 0 hexa를 제외한 모든 코드와 결합이 가능하다.

표 6. 원천 부호화 규칙 ; 4 x 4비트[1]
Table 6. Source coding rule ; 4 x 4-bits

hexa	상위 비트열	하위 비트열	
		조합제한	조합가능
O	OOOO	O,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F	X
1	OOO1	O	1,2,3,4,5,6,7,8,9,A,B,C,D,E,F
2	OO1O	O, 1	2,3,4,5,6,7,8,9,A,B,C,D,E,F
3	OO11	O	1,2,3,4,5,6,7,8,9,A,B,C,D,,E,F
4	O1OO	O,1,2,3	4,5,6,7,8,9,A,B,C,D,E,F
5	O1O1	O	1,2,3,4,5,6,7,8,9,A,B,C,D,E,F
6	O11O	O, 1	2,3,4,5,6,7,8,9,A,B,C,D,E,F
7	O111	O	1,2,3,4,5,6,7,8,9,A,B,C,D,E,F
8	1OOO	O,1,2,34,5,6,7	8,9,A,B,C,D,E,F
9	1OO1	O	1,2,3,4,5,6,7,8,9,A,B,C,D,E,F
A	1O1O	O, 1	2,3,4,5,6,7,8,9,A,B,C,D,E,F
B	1O11	O	1,2,3,4,5,6,7,8,9,A,B,C,D,E,F
C	11OO	O, 1, 2, 3	4,5,6,7,8,9,A,B,C,D,E,F
D	11O1	O	1,2,3,4,5,6,7,8,9,A,B,C,D,E,F
E	111O	O, 1	2,3,4,5,6,7,8,9,A,B,C,D,E,F
F	1111	O	1,2,3,4,5,6,7,8,9,A,B,C,D,E,F

표 7은 두 개의 문자코드가 결합될 때를 위한 조합규칙을 나타낸 것으로 전송되는 문자의 선행 바이트의 최상위 비트열과 다음에 전송되는 문자의 최하위 비트열과 결합이 발생한다. 그러므로 첫 번째 바이트의 최상위 비트열이 어떻게 구성되는가에 따라 다음 비트열과의 조합에 의해 4개 이상의 연속“0”의 비트를 발생시킬 수 있다. 예를 들어 표 7에서 최상위 비트열이 “0001”일 때 그에 이어지는 비트열에 (0, 2, 4, 6, 8, 10, 12, 14)hexa 중 하나가 있을 경우에 4개 이상의 연속“0”의 비트열이 발생된다[1]

표 7. 8비트(1바이트간) 조합규칙[1]
Table 7. Composition rule between hexa bytes

16진수	첫번째 바이트 최상위 비트열	두 번째 바이트 최하위비트열	
		조합제한	조합가능
0	0000	모두	-
1	0001	0,2,4,6,8,10,12,14	1,3,5,7,9,11,13,15
2	0010	0, 4, 10, 14	1,2,3,5,6,7,8,9,11,12,13,15
3	0011	0, 4, 10, 14	1,2,3,5,6,7,8,9,11,12,13,16
4	0100	0, 8	1,2,3,4,5,6,7,9,10,11,12,13,14,15
5	0101	0, 8	1,2,3,4,5,6,7,9,10,11,12,13,14,15
6	0110	0, 8	1,2,3,4,5,6,7,9,10,11,12,13,14,15
7	0111	0, 8	1,2,3,4,5,6,7,9,10,11,12,13,14,15
8	1000	0	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
9	1001	0	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
A	1010	0	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
B	1011	0	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
C	1100	0	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
D	1101	0	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
E	1110	0	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
F	1111	0	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15

4.2 원천부호화규칙과 라틴어 부호 체계

표 2의 유니코드 1.0 라틴어 부호 체계는 표1의 BMP부분에 해당된다. 즉, 부호 체계의 4바이트 중 상위 2바이트가 (0000)hexa로 구성된다. 그리고 하위 2바이트는 (0000)hexa~(007F)hexa로 구성되어 있다. 즉, 하위 2바이트중 상위1바이트가 (00)hexa로 구성된다. 만약에 이 부호 체계를 통신용으로 사용한다고 한다면, 표 6과 표 7의 원천부호화규칙에서 최악의 부호 체계에 해당된다. 또한 하위2바이트중 하위1바이트의 구성도 원천 부호화 규칙을 적용할 경우 다음과 같은 부적합한 점이 발견된다. 첫째 라틴어 문자의 경우에는 대문자 A, B, C와 P, 소문자 a와 p가 조합규칙에 위배되는 것으로 나타났다. 숫자의 경우에는 0이 규칙에 위배된다. 기호의 경우에는 !가 규칙에 위배된다. 통신프로토콜 문자 SOH, STX, ETX, EOT, ENQ, ACK, BEL 등이 규칙에 위배되는 것으로 나타났다.

4.3 유니코드 라틴어 부호 체계 개선방안

본 논문은 다음과 같은 두 가지의 경우에 대하여 개선방안을 제시하였다. 첫째는 유니코드 1.0의 라틴어 부호 체계를 그대로 통신에 사용할 경우이다. 두 번째는 유니코드 1.0의 라틴어 부호 체계가 UTF-8 코드로 변환하여 통신용으로 사용할 경우에 대한 것이다.

가. 유니코드를 통신용 부호로 사용할 경우

1) 원천부호화 적용시 문제점

다음은 현재의 유니코드 3.0의 라틴어 부호 체계를 통신용 부호로 사용할 경우에 주는 영향을 분석한 것이다. 유니코드 부호 체계 내에서 연속 4개의 비트 0이 발생하는 16진수의 개수를 보여주는 것이다. 즉, 이것들은 표 6의 원천 부호화 규칙에 부합되지 않는 부호의 개수이다.

O 현재 부호 체계 내 연속 4개 0비트 hexa 개수

다음은 현재 부호 체계내에서 4개의 연속 0의 비트, 즉 원천 부호화 규칙에 부합되지 않는 부호의 개수를 보여 주는 것이다.

- 하위 문자판(2 바이트)

4개짜리((0000)) 1개 = 1개
3개짜리(000x) 15개 = 15개
2개짜리(001x) 15개 * 7열 = 105개
1개짜리(xx10) = 7개
총 138개

- 상위 문자판 (2 바이트)

4개짜리(0000) * 16행 * 8열 = 512개

- 누적 총계 = 650개

다음은 현재 부호 체계 내에서 라틴어에 대한 것만 4개의 연속 0의 비트, 즉 원천 부호화 규칙에 부합되지 않는 부호의 개수를 보여 주는 것이다.

- 상위문자판 (2 바이트)

4개짜리 (0000); 4개 * 라틴어 52개 = 208개

- 하위문자판(2 바이트)

2개짜리(00xx); 2개*라틴어 52개 = 104개
1개짜리(xxx0) = 2개

- 누적 총계 = 314개

2) 개선방안

가) 방안 1 : 현재 부호 체계 수용

첫 번째 개선방안은 표 8에서 보여주는 것과 같은 부호 체계이다. 표 8은 표 3의 상위문자판과 하위문자판을 바꾸고 하위문자판의 부호배열체계는 표 3의 현행과 같이 유지하는 방안이다. 즉 상위문자판은 현행 0000를 9111로 바꾸었고 하위문자판은 0000-0070을 9110-9180으로 바꾸었다. 그러므로 이 개선방안은 하위문자판의 부호배열체계에 대하여는 [1]에서 제시한 원천 부호화 규칙을 적용하지 않았다는 단점이 있다. 그러나 현재의 부호배열체계를 그대로 수용한다는 측면에서 부호 체계개선으로 인한 혼란을 감소시킬 수 있는 장점이 있다.

나) 방안 2 : 사용빈도 기준, 문자간 재배치

두 번째 개선방안은 표 9와 같이 각 문자판 열내에서 사용빈도에 따라 문자를 재배열하는 방안이다. 표 9의 DC1은 장치제어부호이다. ESC와 DEL은 컴퓨터 자체 제어신호로서 네트워크상에 전송되는 신호가 아니다. 라틴어 x, y, z는 918x문자판 내에서 다른 문자에 비하여 상대적으로 사용빈도가 매우 낮은 문자이다. 916x 문자판내에 있는 P, Q는 사용빈도가 높은

편에 속하는 문자이다. 따라서 동일 문자판 내에서 상대적으로 네트워크로 전송되는 빈도가 낮은 기호로 대치하였다.

다) 방안3 : 사용빈도 기준, 문자열간 대치

세 번째 방안으로는 표 10과 같이 문자열 간에 상호 대치하는 방안이다. 즉, 전송빈도가 라틴어 문자보다 상대적으로 낮은 913x의 기호 문자열을 라틴어 문자열인 918x 문자열과 대치시켰다. 다음은 숫자 문자열인 914x와 통신제어 신호 문자열인 912x 문자열과 대치시켰다. 한편 이러한 방법으로 할 경우 사용빈도가 높은 라틴어 P와 p 그리고 아라비아 숫자 0과 2 및 통신장치 제어문자가 원천 부호화 규칙에 부합되지 않는 단점이 있다.

라) 방안4 : 문자 및 문자열 대치

네 번째 개선방안은 두 번째 방안과 세 번째 방안을 혼합하는 방안이다. 표 11은 이 방안을 보여 주는 것이다.

마) 방안5 : 라틴어 부호열만 재배치

다섯 번째 개선방안은 라틴어 부호열에 대하여만 원천 부호화 규칙에 부합되도록 하는 것이다. 표 12는 이것을 보여 주는 것이다. 즉 916x 문자열과 918x 문자열 내에서 만 조정하는 것이다. 916x 문자열의 경우에는 라틴어 대문자 P, Q가 이에 해당된다. 918x 문자열의 경우에는 p, q, r, s, t, u, v, w가 이에 해당된다. 이 방식은 표 9에서 제시한 개선방안의 일부를 도입한 방식이다.

3) 개선방안별 비교

표 13은 이 다섯 가지 개선방안을 비교하기 위해 각 개선방안에서 원천 부호화 규칙에 부합되지 않는 문자를 비교한 것이다. 이 표 13에서 보는 바와 같이 표 8의 개선방안이 원천 부호화 규칙에 맞지 않는 라틴어와 숫자를 가장 많이 가지고 있다. 표 9, 표 10과 표 12의 개선방안의 경우에는 라틴어가 각각 세 개씩 있지만, 사용빈도수를 볼 때 표 9와 표 12의 개선방안이 상대적으로 전송효율측면에서 유리함을 알 수 있다. 표 11의 네 번째 개선방안에는 연속 4개의 0비트를 갖는 라틴어나 숫자가 없음을 알 수 있다. 그러므로 전송효율 측면에서는 표 11의 네 번째 개선방안이 가장 좋은 방안으로 나타났다. 그러나 현재 표준부호체계인 표 3의 부호 체계와의 일치성 관점에서 볼 때는 표 8의 첫 번째 개선방안이 가장 부합되고 다음으로는 표 12의 개선방안임을 알 수 있다. 본 논문에서는 전송효율관점에서 라틴어 부호 체계에 대한 개선방안을 제시하는 것임으로 표 11의 네 번째 개선방안을 라틴어 부호 체계에 대한 최적방안으로 제시하였다.

표 13. 각 부호 체계 개선방안 비교 : 원천 부호화 규칙에 부합되지 않는 문자

Table 13. Compare of each proposed way : unmatching characters with source coding rule

방안	원천 부호화 규칙에 위배되는 문자
방안1	NUL DLE DC1 SP 0 1 2 3 @ P Q p q r s t u v w
방안2	NUL DLE ESC SP DEL @ < = > ? ^ _ { | } ~ x y z
방안3	NUL DLE DC1 DC2 DC3 SP 0 1 @ ! " # $ % & ' p P Q
방안4	NUL DLE ESC DC1 DC2 SP @ < = { ^ _ ! " # $ % & '
방안5	NUL DLE DC1 DEL SP 0 1 2 3 @ ^ _ { | } ~ x y z

4) 현재 부호 체계와 개선방안 부호 체계별 전송효율비교분석

가) 개선방안별 연속 4개 0비트 hexa 개수

다음은 본 논문에서 제시한 다섯 개의 각 개선방안별 부호 체계내에서 4개의 연속 0의 비트가 발생하는 개수를 산출한 것이다.

O 방안1(표 8): 전체 20개, 라틴어 10개
O 방안2(표 9): 전체 20개, 라틴어 3개
O 방안3(표 10):전체 20개, 라틴어 3개
O 방안4(표 11):전체 20개, 라틴어 0개
O 방안5(표 12):전체 20개, 라틴어 3개

나) 현재 부호 체계와 개선방안별 효율비교

O 개선방안 1의 경우 효율개선효과
- 부호 체계 전체 문자대비 개선효율 :
= 현재 부호 체계 전체 문자중 연속 4개 0

비트 16진수 개수/개선방안1 부호 체계내 연속 4개 0비트 16진수 개수
= 729/20
= 36.45배
- 부호 체계 내 라틴어 문자대비 개선효율
= 현재 부호 체계 내 라틴어 중 연속 4개 0 비트 16진수 개수/개선방안1 부호 체계 내 라틴어중 연속4개 0비트 16진수 갯수
= 314/10
= 31.40배

O 개선방안 2, 3, 5의 경우 효율개선효과
- 부호 체계 전체 문자대비 개선효율 :
= 현재 부호 체계 내 전체 문자중 연속 4개 0비트 16진수 개수/개선방안 2, 3, 5 부호 체계내 연속 4개 0비트 16진수개수
= 729/20
= 36.45배
- 부호 체계 내 라틴어 문자대비 개선효율
= 현재 부호 체계내 라틴어 중 연속 4개 0 비트 16진수 개수/개선방안 2, 3, 5 부호 체계 내 라틴어 중 연속 4개 0비트 16진수 개수
= 314/3
= 104.5배

O 개선방안 4의 경우 효율개선효과
- 부호 체계 전체 문자대비 개선효율 :
= 현재 부호 체계 내 전체 문자 중 연속 4개 0비트 16진수 개수/개선방안4 부호 체계내 연속 4개 0비트 16진수 개수
= 729/20
= 36.45배
- 부호 체계 내 라틴어 문자대비 개선효율
= 현재 부호 체계 내 라틴어 중 연속 4개 0비트 16진수 개수/개선방안 4 부호체계내 라틴어 중 연속 4개 0비트 16진수 개수
= 314/0
= 314.0배

나. UTF-8 코드체계를 고려한 개선 방안

1) 개선방안

유니코드 1.0의 라틴어 부호 체계는 UTF-8 코드로 변환될 경우 유니코드의 하위문자판을 구성하고 있는 2바이트 중에서 하위 1바이트만 사용한다. 표 14는 각 유니코드를 UTF코드로 변환하는 규칙을 보여 주는 것이다. 특히 표 14의 U+007F에 대한 변환은 본 논문의 연구대상인 라틴어가 포함되어 있는 코드체계에 해당된다.

표 14. 유니코드 UTF-8 변환방법
Table 14. Method for converting unicode to UTF-8

유니코드 최종코드점	UTF-8 코드; x는 유니코드 구성 각 비트
U+007F	0xxxxxxx
U+07FF	110xxxxx 10xxxxxx
U+FFFF	1110xxxx10xxxxxx10xxxxxx
U+1FFFFF	11110xxx10xxxxxx10xxxxxx 10xxxxxx
U+3FFFFFF	111110xx10xxxxxx10xxxxxx 10xxxxxx10xxxxxx
U+7FFFFFFF	1111110x10xxxxxx10xxxxxx10xxxxxx10xxxxxx10xxxxxx

그러므로 표 3의 부호 체계에서 상위 문자판을 구성하고 있는 2바이트(0000 hexa)와 하위문자판을 구성하고 있는 2바이트 중, 상위 1바이트(00 hexa)를 제외한 나머지 1바이트로만 구성된다. 표 15~22는 이것을 보여주는 것이다. 표 15와 표 16에서 백색 부분의 코드가 원천 부호화 규칙에 부합되지 않는 코드들이다. 표 17은 표 15와 표 16을 통합한 것이다.

본 논문에서는 표 17을 기준으로 하여 UTF-8의 라틴어 부호 체계에 대한 개선방안을 제시하였다.

가) 방안 1 (표 18) : 각 행내에서 코드간 교체

제1안은 각 행내에서 전송빈도가 높은 문자를 원천 부호화 규칙에 부합되는 코드에 부여하는 것이다. 예를 들어 U+0행내에서는 U+00(NUL)외에 모든 문자의 조건이 동일하므로 현재 상태를 유지한다.

U+1행의 경우에는 통신용 코드인 SYN(U+16)을 정보기기 내부 코드인 ESC(U+1B)와 교체한다.

U+2행의 경우에는 모두 기호문자이므로 현재 상태를 유지한다. U+3행의 경우에는 상대적으로 사용빈도가 낮은 U+3B(;)를 사용빈도가 높은 숫자 0(U+30)과 교체한다. U+4행의 경우에는 라틴어를 중심으로 하여 A로부터 O까지 내에서 가장 사용빈도가 낮은 라틴어를 현재의 A, B, C, H(U+41, 42, 43, 48)와 교체한다. 즉 이 행에서 빈도가 낮은 순위로 정리하면 J, K, B, G, F, M, C, D, L, H, I, N, O, A, E가 된다. 그러므로 A, B. C, H를 B, F, J, K로 교체한다. U+5행의 경우에는 U+50과 U+58의 P와 X를 상대적으로 사용빈도가 낮은 Q, Z와 교체한다.

U+6행의 경우에는 U+61과 U+68의 a와 h를 이 행에서 가장 사용빈도가 낮은 j, k와 교체한다. U+7행의 경우에는 U+70과 U+78의 p와 x를 상대적으로 사용빈도가 낮은 q, z와 교체한다.

나) 방안 2 (표 19): 각 행내 코드간 교체 후 라틴어 차례대로 배열

제2안은 제1안과 같이 각 행내에서 전송빈도가 높은 문자에 원천 부호화 규칙에 부합되는 코드를 부여하되, 라틴어문자의 경우에는 각 행내에서 라틴어 순서대로 재배열하는 것이다.

다) 방안 3 (표 20): 라틴어 대소문자별 행과 열 코드 대치

제3안은 라틴어문자열에 대하여만 원천 부호화 규칙을 적용하되 대문자와 소문자를 각각 하나의 그룹으로 하여 원천 부호화 규칙을 적용한다.

즉, 라틴어 대문자는 U+4행과 U+5행을 하나의 그룹으로 한다. 라틴어 A로부터 Z까지 내에서 가장 사용빈도가 낮은 라틴어를 현재의 A, B, C, H, P, X(U+41, 42, 43, 48, 50, 58)에 배치한다. 라틴어에서 사용빈도가 가장 낮은 순서대로 여섯 개를 정리하면 Z, J, Q, X, K와 V가 된다. 그러므로 A, B. C, H, P, X를 Z, J, Q, X, K와 V로 교체한다. 라틴어 소문자는 U+6행과 7행에 배치되어 있다. 이 두 행에서는 U+61과 U+68의 a와 h, U+70과 U+78의 p와 x를 사용빈도가 가장 낮은 Z, J, Q와 X로 교체한다.

라) 방안 4 (표 21) : 라틴어 대소문자별 코드 대치 후 차례대로 라틴어 배열

제4안은 제3안과 같이 하되 라틴어에 대하여 원천 부화를 적용한 후에 라틴어를 순서대로 재배열하는 것이다.

마) 방안 5 (표 22): 원천 부호화 규칙에 부합되는 코드중심으로 전면 재배열

제5안은 라틴어, 숫자 및 통신용코드를 중심으로 하여 전면 수정한 안이다. 라틴어의 경우 사용빈도가 낮은 Y와 Z 두개의 문자 외에는 모두 원천 부호화 규칙내에 수용되도록 하였다. 그리고 기기 내부제어 코드와 사용빈도가 상대적으로 낮은 기호 등은 원천 부호화 규칙에 부합되지 않는 코드에 배치하였다. 나머지 코드에 배치하였다.

2) 개선방안별 비교

표 23~25은 이상에서 제시한 다섯 개의 개선방안을 비교한 것이다. 표 23은 각 개선방안에서 원천 부호화 규칙에 부합되는 문자의 개수를 비교한 것이다. 이 표 23에서 보여주는 바와 같이 방안 1~4는 현행과 별 차이가 없음을 알 수 있다. 그러나 방안5와는 큰 차이가 있음을 알 수 있다. 즉, 방안 5로 할 경우 라틴어에 대한 개선효과는 96%이상으로 나타났다.

표 23. 유니코드 1.0 C0 제어와 기본 라틴어 부호 체계 내 문자유형별 개수
Table 23. Number of chracters in C0 controls and basic latin code system in unicode 1.0

구분	라틴어	기호	통신	기기 내부	숫자
전체 문자수	52	32	12	22	10
전체 문자 중 원천 부호화 규칙에 부합되는 문자수					
현행	42	25	2	7	8
1~4안	42	25	3	6	8
5안	50	13	11	0	10

표 24는 라틴어에 대하여 현행과 다섯 개의 각 방안별 사용빈도에 따른 비교표이다. 이 표에서 보는 바

와 같이 현행의 경우 사용빈도가 42.6이고 제1안과 제2안은 8.73, 제3안과 제4안은 4.93 그리고 제5안은 2.46이다. 이것은 원천 부호화 규칙에 부합되지 않는 값으로 큰 값일수록 부적합도가 큰 것이다.

표 25는 현행 유니코드 UTF-8 부호 체계 대비 각 방안별 개선도를 나타내는 것이다. 현행대비 제1안과 제2안의 경우에는 4.88배, 제3안과 제4안의 경우에는 8.64배 그리고 제5안의 경우에는 17.3배 개선되는 것으로 나타났다.

표 24. 유니코드 1.0 C0 제어와 기본 라틴어 부호 체계내 및 각 방안별 원천 부호화 규칙에 어긋나는 라틴어 부호

Table 24. Unmatched character codes in C0 controls and basic latin code system in unicode 1.0 and in proposed methods

현행		제1안		제2안		제3안		제4안		제5안	
A	9.43	B	2.1	B	2.1	J	0.3	J	0.3	Y	2.25
a	9.43	F	2.61	F	2.61	j	0.3	j	0.3	Z	0.21
B	2.1	J	0.3	J	0.3	K	1.15	K	1.15		
C	4.09	J	0.3	j	0.3	Q	0.35	Q	0.35		
H	5.72	K	1.15	K	1.15	q	0.35	q	0.35		
h	5.72	k	1.15	k	1.15	V	1.34	V	1.34		
P	2.69	Q	0.35	Q	0.35	X	0.36	X	0.36		
p	2.69	q	0.35	q	0.35	x	0.36	x	0.36		
X	0.36	Z	0.21	Z	0.21	Z	0.21	Z	0.21		
x	0.36	z	0.21	z	0.21	z	0.21	z	0.21		
계	42.6	계	8.73	계	8.73	계	4.93	계	4.93	계	2.46

표 25. 현행 부호 체계대 제안방안 부호 체계별 개선도

Table 25. Improvement standards of each of proposed methods

개선방안	1, 2안	3, 4안	5안
개선도	4.88	8.64	17.3

V. 결 론

본 논문은 데이터전송에 대한 효율적 측면에서 국제표준 문자부호 체계인 유니코드(Unicode) 3.0 부호 체계에 포함된 라틴어 문자의 부호 체계에 대하여 연구하였다. 유니코드는 4바이트 32비트 구조를 가지고 있다. 유니코드에 의하여 원천 부호화된 문자는 데이터 전송로로 전송될 때 UTF 등의 전송부호 체계로 변환된다. 특히 UTF의 경우에는 유니코드를 기반으로 하여 변환된다. 본 본문은 유니코드 3.0의 라틴어 부호 체계가 데이터통신의 전송효율 측면에서 적정한지 여부를 연구한 결과 적정치 않은 것으로 나타났다. 따라서 그 결과에 따라 HDB-3 방식에 적용하기 위해 [1][19][20]에서 제시된 원천 부호화 규칙을 적용하여 개선방안을 제시하였다.

그 결과 본 논문에서 제시한 개선된 유니코드 라틴어 부호 체계와 UTF-8 라틴어 부호 체계를 적용할 경우, 회선부호기의 스크램블러 운용효율을 유니코드를 통신용으로 사용할 경우 최소 36.45배에서 최대314배, 그리고 제시된 UTF-8 라틴어 부호 체계를 적용할 경우 최대 17배에서 최소 4.8배까지 개선되는 것으로 나타났다.

표 8. 유니코드 1.0 라틴어 부호 체계 개선안 ; 상위 및 하위 문자판 변경, 현행 문자배열체계유지

Table 8. New proposal C0 controls and basic latin code system in unicode 1.0 system ; upper and lower plane change and maintain the current characters array

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
911	**NUL**	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
912	**DLE**	**DC1**	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
913	**SP**	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
914	**0**	**1**	**2**	**3**	4	5	6	7	8	9	:	;	<	=	>	?
915	**@**	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
916	**P**	**Q**	R	S	T	U	V	W	X	Y	Z	[	₩	]	^	_
917		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
918	**p**	**q**	**r**	**s**	**t**	**u**	**v**	**w**	x	y	z	{	|	}	~	DEL

본 논문에서는 본 논문에서 제시된 유니코드 라틴어 부호 체계와 현재의 유니코드 라틴어 부호 체계와의 호환성에 대한 것은 논외로 하였다. 그러므로 향후 이 부분에 대한 연구는 본 논문에서 제시한 연구결과가 통신현장에서 활용되도록 하는데 큰 영향을 미칠 것으로 판단된다.

표 9. 유니코드 1.0 라틴어 부호 체계 개선안 ; 각 행내에서 원천 부호화 규칙에 따라 부호배열
Table 9. New proposal C0 controls and basic latin code system in unicode 1.0 system ; codes array in each row with rule of source coding

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
911	**NUL**	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
912	**DLE**	ESC	***DC1***	***DC2***	***DC3***	***DC4***	***NAK***	***SYN***	***ETB***	***CAN***	***EM***	***SUB***	FS	GS	RS	US
913	**SP**	!	"	#	$	%	&	'	(	)	*	+	,	−	.	/
914	<	=	>	?	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***:***	***;***
915	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
916	^	_	***P***	***Q***	***R***	***S***	***T***	***U***	***V***	***W***	***X***	***Y***	***Z***	[	₩	]
917		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
918	{	|	}	~	DEL	x	y	z	***p***	***q***	***r***	***s***	***t***	***u***	***v***	***w***

표 10. 유니코드 1.0 라틴어 부호 체계 개선안 ; 각 행간 원천 부호화 규칙에 따라 부호배열
Table 10. New proposal C0 controls and basic latin code system in unicode 1.0 system ; codes array between each row with rule of source coding

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
911	***NUL***	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
912	***0***	***1***	***2***	***3***	4	5	6	7	8	9	:	;	<	=	>	?
913	***p***	***q***	***r***	***s***	***t***	***u***	***v***	***w***	x	y	z	{	|	}	~	DEL
914	***DLE***	***DC1***	***DC2***	***DC3***	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
915	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
916	***P***	***Q***	R	S	T	U	V	W	X	Y	Z	[	₩	]	^	_
917		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
918	**SP**	**!**	″	***#***	***$***	***%***	***&***	′	(	)	*	+	,	−	.	/

표 11. 유니코드 1.0 라틴어 부호 체계 개선안 ; 각 행내 및 행간 원천 부호화 규칙에 따라 부호배열
Table 11. New proposal C0 controls and basic latin code system in unicode 1.0 system ; code array in row and between row with rule of source coding

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
911	***NUL***	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
912	<	=	>	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***:***	***;***	?
913	{	|	}	~	DEL	x	y	z	***p***	***q***	***r***	***s***	***t***	***u***	***v***	***w***
914	**DLE**	ESC	***DC1***	***DC2***	***DC3***	***DC4***	***NAK***	***SYN***	***ETB***	***CAN***	***EM***	***SUB***	FS	GS	RS	US
915	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
916	^	_	***P***	***Q***	***R***	***S***	***T***	***U***	***V***	***W***	***X***	***Y***	***Z***	[	₩	]
917		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
918	**SP**	**!**	″	***#***	***$***	***%***	***&***	′	(	)	*	+	,	−	.	/

표 12. 유니코드 1.0 라틴어 부호 체계 개선안 ; 라틴어 부호 체계만 원천 부호화 규칙에 따라 부호배열
Table 12. New proposal code system of C0 controls and basic latin code system in unicode 1.0 system ; code array only latin characters

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
911	***NUL***	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
912	***DLE***	***DC1***	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
913	**SP**	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
914	***0***	***1***	***2***	***3***	4	5	6	7	8	9	:	;	<	=	>	?
915	**@**	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
916	^	_	***P***	***Q***	***R***	***S***	***T***	***U***	***V***	***W***	***X***	***Y***	***Z***	[	₩	]
917		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
918	{	|	}	~	DEL	x	y	z	***p***	***q***	***r***	***s***	***t***	***u***	***v***	***w***

표 15. UTF-8 부호 체계내 C0제어와 기본 라틴어 원천 부호화 규칙 부합 부호(표 6)
Table 15. C0 controls and basic latin code system in unicode 1.0 system(Table 6)

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	₩	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

표 16. UTF-8 부호 체계내 C0제어와 기본 라틴어 원천 부호화 규칙 부합 부호(표 7)
Table 16. C0 controls and basic latin code system in unicode 1.0 system (Table 7)

상하	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	₩	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

* 채색부분이 원천 부호화 규칙에 부합되는 부호 (이하 모든 그림도 동일함)

표 17. UTF-8 부호 체계내 C0제어와 기본 라틴어 원천 부호화 규칙 부합 부호 ; 표 6과 표7을 합친 것
Table 17. C0 controls and basic Latin code system in ynicode 1.0 system ; combine between Table 6 and Table 7

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	₩	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

표 18. UTF-8 라틴어 부호 체계 개선안 ; 각 행내에서 원천 부호화 규칙에 따라 부호배열
Table 18. New proposal C0 controls and basic latin code system in UTF-8 code system ; codes array in each row with rule of source coding

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	ESC	ETB	CAN	EM	SUB	SYN	FS	GS	RS	US
2	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	;	1	2	3	4	5	6	7	8	9	:	0	<	=	>	?
4	@	B	F	J	D	E	A	G	K	I	C	H	L	M	N	O
5	Q	P	R	S	T	U	V	W	Z	Y	X	[	₩	]	^	_
6	`	j	b	c	d	e	f	g	k	i	a	h	l	m	n	o
7	q	p	r	s	t	u	v	w	z	y	x	{	|	}	~	DEL

표 19. UTF-8 라틴어 부호 체계 개선안 ; 각 행내에서 원천 부호화 규칙에 따라 교체하고 라틴어 순으로 부호 배열
Table 19. New proposal C0 controls and basic latin code system in UTF-8 code system ; codes array in each row with rule of source coding and orderly rearrange latin characters

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	ESC	ETB	CAN	EM	SUB	SYN	FS	GS	RS	US
2	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	;	1	2	3	4	5	6	7	8	9	:	0	<	=	>	?
4	@	B	F	J	A	C	D	E	K	G	H	I	L	M	N	O
5	Q	P	R	S	T	U	V	W	Z	X	Y	[	₩	]	^	_
6	`	j	a	b	c	d	e	f	k	g	h	i	l	m	n	o
7	q	p	r	s	t	u	v	w	z	x	y	{	|	}	~	DEL

표 20. UTF-8 라틴어 부호 체계 개선안 ; 라틴어 대소문자 행내에서 원천 부호화 규칙에 따라 부호배열
Table 20. New proposal C0 controls and basic latin code system in UTF-8 code system ; array latin characters separately between big letter and small letter of latin

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	J	K	Q	D	E	F	G	V	I	A	B	L	M	N	O
5	X	Q	R	S	T	U	H	W	Z	Y	P	[	₩	]	^	_
6	`	j	b	c	d	e	f	g	q	i	a	k	l	m	n	o
7	x	h	r	s	t	u	v	w	z	y	p	{	|	}	~	DEL

표 21. UTF-8 라틴어 부호 체계 개선안 ; 라틴어 대소문자 행내에서 원천 부호화 규칙에 따라 부호배열 후 라틴어 순서대로 부호 재배열
Table 21. New proposal C0 controls and basic latin code system in UTF-8 code system ; array latin characters separately between big letter and small letter of latin and then orderly rearrange latin characters

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	J	K	Q	A	B	C	D	V	E	F	G	H	I	L	M
5	X	N	O	P	R	S	T	U	Z	W	Y	[	₩	]	^	_
6	`	j	a	b	c	d	e	f	q	g	h	i	k	l	m	n
7	x	o	p	r	s	t	u	v	z	w	y	{	|	}	~	DEL

표 22. UTF-8 라틴어 부호 체계 개선안 ; 라틴어 중심 원천 부호화 규칙 적용
Table 22. New proposal C0 controls and basic latin code system in UTF-8 code system ; latin characters-oriented code array with the source coding rule

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SUB	ESC	FS	GS	RS	US	EM	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	!	"	$	#	%	&	(	'	)	*	,	+	.	-	/
2	SP	SOH	STX	ETX	^	EOT	ENQ	ACK	NAK	SYN	|	ETB	CAN	BEL	}	_
3	DEL	@	0	1	=	2	3	4	5	6	`	7	8	9	{	~
4	:	[	Z	Y	X	W	V	U	DC1	T	S	R	Q	P	O	N
5	;	₩	A	B	C	D	E	F	DC2	G	H	I	J	K	L	M
6	<	]	z	y	x	w	v	u	DC3	t	s	r	q	p	o	n
7	>	?	a	b	c	d	e	f	DC4	g	h	i	j	k	l	m

감사의 글

본 논문은 한세대학교의 연구지원사업에 의해 수행되었습니다.

참고 문헌

[1] 홍완표, “데이터 전송 효율을 고려한 4비트행x4비트열 2 바이트 문자 부호화 규칙에 관한 연구”, 한국항행학회논문지, 제15권, 제5호, pp. 749-756, 10월, 2011년.
[2] American Standards Association, "American Code (July 6, 1999). for Information Interchange", ASA X3.4- 1963, 17 June, 1963.
[3] American National Standards Institute, "American National Standard for Information Systems-Coded Character Sets 7-Bit American National Standard Code for Information Interchange (7-Bit ASCII)", ANSI X3.4-1986, Inc., 26 March 1986.
[4] RFC 20 “ASCII format for Network Interchange” October 1969 (http://tools.ietf.org/ html/rfc20)
[5] http://en.wikipedia.org/wiki/EBCDIC
[6] http://en.wikipedia.org/wiki/Unicode
[7] 산업자원부 기술표준원,"정보 교환용 부호계 (한글 및 한자) 부속서 3. 보조 부호계(2바이트 조합형 부호계", KS X 1001 : 2004. 2004년 12월 28일 개정.
[8] 산업자원부 기술표준원,“KS C 5601 : 1987 (1987년 고침) : 정보 교환용 부호(한글 및 한자)”, 2004년 12월 28일 개정.
[9] ITU-T Recommendation G.703, "Physical/ electrical characteristics of hierarchical digital interfaces" pp. 24-41, Oct. 1998.
[10] TTA Standard, “Test Method for Telecommunication Terminal Equipment” TTAS. KO-05.0028/R1, pp306-451, Revised on 23 Dec. 2004.
[11] Behrouz A. Forouzan, “Data communications" McGraw Hill Korea, pp. 132-134. 2008.
[12] Behrouz A. Forouzan, “Data communications" McGraw Hill Korea, p.1031. 2007.
[13] Behrouz A. Forouzan, “Data communications" McGraw Hill Korea, p1032. 2007.
[14] http://en.wikipedia.org/wiki/Unicode_plane
[15] http://en.wikipedia.org/wiki/Unicode#Versions
[16] http://100.naver.com/100.nhn?docid=780033, 위기언어, 네이버 백과사전
[17] http://oxforddictionaries.com/words/what-is-the-frequency-of-the-letters-of-the-alphabet-in-english
[18] http://www.world-english.org/english500.htm
[19] 홍완표,“데이터 전송효율을 고려한 3x4비트 1바이트 문자부호화 규칙에 관한 연구” 한국전자통신학회논문지, 6권, 4호, pp. 499-504, 08, 2011.
[20] 홍완표,“데이터통신 전송효율과 ASCII 부호 체계 고찰” 한국전자통신학회논문지, 6권, 5호, pp. 657-664, 10, 2011.

저자 소개

홍완표(Wan-Pyo Hong)

1991년 서울과학기술대학교 전자공학과 졸업(공학사)
1994년 연세대학교대학교 공학대학원 산업공학과 졸업(공학석사)
1999년 광운대학교 대학원 전자공학과 졸업(공학박사)
1990년 전기통신기술사합격
1991년 정보통신부 5급특별채용고시합격 본부 통신정책실, 전파방송관리국, 정보화기획실
1997년 삼성전자(주) 통신사업부 전송영업그룹장
1999년 광운대학교 연구전담교수
2000년 한국정보통신기술협회장
2002년 한세대학교 IT학부 정보통신공학전공 교수
한세대학교 정보통신연구소장
※ 관심분야 : 위성통신방송, 문자코딩, 통신정책