Regular paper http://dx.doi.org/10.13067/JKIECS.2013.8.7.1011

AMI/HDB-3 회선부호화와 한·중·일 한자 유니코드 체계 고찰

태동진* · 홍완표**

Consideration of CJK Joint Hanja Unicode when is used in AMI/HDB-3 Line Coding

Dong-Zhen Tai* · Wan Pyo Hong**

요 약

본 논문은 한중일통합 한자 유니코드 부호 집합체계가 원천부호화규칙에 위배되는 정도를 분석하였다. 본 연구에서는 한중일통합 한자의 유니코드 중에서 사용빈도 수가 높은 문자 150개를 대상으로 하여 연구하였다. 이 한중일통합 한자 150개 문자의 사용 빈도율은 한중일통합 한자 유니코드 전체 사용빈도율의 약 50%에 해당된다. 본 연구에서는 한중일통합 한자 유니코드를 AMI회선부호화 방식과 HDB-3 스크램블링 방식을 사용할 경우를 대상으로 하였다. 분석결과 150개의 문자중 원천부호화 규칙에 위배되는 문자는 총 77개 였다. 이들 문자들의 사용 빈도율에 의한 원천부호화 규칙 위배율은 약28%였다. 결과적으로 이 원천부호화 규칙에 위배되는 문자들을 사용빈도가 낮고 원천부호화 규칙에 부합되는 문자부호로 대체 할 때, 회선부호기에서의 회선부호 처리율을 약37%만큼 개선시킬 수 있음을 나타냈다.

ABSTRACT

This paper analyses the violation rate of CJK joint Chines character Unicode to the source code rule. In the paper, Chinese character 150ea in Chinese Unicode which have relatively a higher frequency in use of a character was chosen to study. The frequency rate in use of the 150ea characters is about 50% of the total frequency rate of the Chinese characters. The study was applied the AMI/HDB-3 line coding/scrambling and HDLC protocol, According to the analyses, the number of violated characters were 77ea of 150 ea, frequency rate in use 29%. Therefore, when the violated 77ea characters are replaced to the matched character codes to the source coding rule, the processing rate of the line coder can be improved about 37%.

키워드

Chinese Character, Line Coding, AMI, HDB-3, HDLC
한자 문자, 회선부호화, AMI, HDB-3, HDLC

Ⅰ. 서 론

정보기기에서 처리되어 통신망으로 전송되는 문자 등의 데이터는 두 단계의 부호화과정을 거친다. 첫 번째 부호화 과정은 OSI (Open system Interoperability) 7계층의 6계층인 표현계층에서 이루어진다. 이 부호화과정은 원천부호화 과정이라고 한다. ASCII [1][2][3], EBCDIC[4], Unicode[5] 및 KS X 1001/1003 등이 이에 해당된다. 두 번째 부호화 과정은 1계층인 물리계층에서 이루어지는데 이 과정을 회선부호화라

* 한세대학교 IT융합과(xiaopang817@naver.com)
** 교신저자(corresponding author) 한세대학교 정보통신공학과(wphong@hansei.ac.kr)
접수일자 : 2013. 05. 06 심사(수정)일자 : 2013. 06. 20 게재확정일자 : 2013. 07. 23

고 한다. 회선부호화 방식에는 NRZ, RZ 및 AMI 등이 있다. 특히 AMI방식은 장거리 정보전송에 적합한 방식이다[6]. 회선부호화방식으로 AMI방식을 사용할 경우 비트 0의 연속발생으로 인한 동기 상실을 방지하기 위해 스크램블링 기술을 사용한다. 대표적인 스크램블링 방식으로는 HDB-3 [6] 와 B8ZS [7] 방식이 있다. 본 논문에서는 연속 4개 이상의 비트 0의 전송을 방지하는 HDB-3방식을 적용하였다. 따라서 원천부호화 과정중 비트 0이 연속4개 이상 존재하는 비트 조합이 많을 경우 회선부호화 과정에서 스크램블링이 많이 발생하며 결과적으로 데이터의 전송효율에 영향을 주게 된다. 한편 OSI의 2계층인 데이터링크 계층에서 HDLC 통신규약을 사용할 때, 프레임내의 문자 비트열중에 FLAG 구성 비트열과 유사한 비트열이 있을 경우에, FLAG비트열로 오인되는 것을 방지하기 위해 비트 또는 문자 스터핑(stuffing)을 하게 된다. 즉 문자 비트열에 이러한 비트열이 많게 되면 통신규약 프레임체계를 변경시키게 되며 데이터의 전송율도 떨어뜨리게 된다.

본 논문에서는 이러한 것을 토대로 하여 한자 유니코드에 대한 원천부호화 위배율이 한자 문자의 사용빈도율을 기준으로 하여 어느 정도되는지 분석하였다. 현재 유니코드 한자부호체계에는 총 20684개의 문자부호가 있다. 본 논문에서는 이 문자 부호 중에서 전체 사용 빈도율이 50%에 해당되는 150자의 문자에 대하여 분석하였다. 본 논문은 참고문헌 [8]에서 제시하고 있는 원천부호화 규칙을 적용하여 분석하였다. 이를 통하여 원천부호화 규칙을 적용할 경우 개선되는 회선부호화 과정에서 얻어지는 개선효과를 정량적으로 산출하였다.

II. 문자원천부호화규칙과 한자 유니코드 부호체계

2.1 문자의 원천부호화규칙

표 1은 참고문헌 [8]에서 제시하고 있는 문자의 원천부호화 규칙이다. 이 규칙은 OSI의 2계층인 데이터링크 계층에 적용되는 HDLC프로토콜과 1계층인 물리계층에서 수행되는 AMI회선부호방식과 HDB-3 스크램블링 기술을 적용한 것이다.

표 1. 문자 부호화 규칙 ; 4 비트 x 4비트[8]
Table 1. Characters coding rule ; 4-bit x 4-bit

HEXA	upper bits	lower bits	
		Unavailable connection	Available connection
0	0000	0,1,2,3,4,5,6,7,8,9,A,B,C,D,E,F	X
1	0001	0,F	1,2,3,4,5,6,7,8,9,A,B,C,D,E
2	0010	0,1	2,3,4,5,6,7,8,9,A,B,C,D,E,F
3	0011	0,E,F	1,2,3,4,5,6,7,8,9,A,B,C,D,
4	0100	0,1,2,3	4,5,6,7,8,9,A,B,C,D,E,F
5	0101	0,F	1,2,3,4,5,6,7,8,9,A,B,C,D,E
6	0110	0,1	2,3,4,5,6,7,8,9,A,B,C,D,E,F
7	0111	0,C,D,E,F	1,2,3,4,5,6,7,8,9,A,B
8	1000	0,1,2,3,4,5,6,7	8,9,A,B,C,D,E,F
9	1001	0,F	1,2,3,4,5,6,7,8,9,A,B,C,D,E
A	1010	0,1	2,3,4,5,6,7,8,9,A,B,C,D,E,F
B	1011	0,E,F	1,2,3,4,5,6,7,8,9,A,B,C,D
C	1100	0,1,2,3	4,5,6,7,8,9,A,B,C,D,E,F
D	1101	0,F	1,2,3,4,5,6,7,8,9,A,B,C,D,E
E	1110	0,1	2,3,4,5,6,7,8,9,A,B,C,D,E,F
F	1111	0,8,9,A,B,C,D,E,F	1,2,3,4,5,6,7

2.2 원천부호화규칙과 사용빈도가 높은 150개 한자 문자 부호

표 2는 Unicode 한중일통합 한자[9]에 규정된 부호집합에서 사용빈도가 높은 순위로 추출한 150개 문자부호[10]와 표 1의 원천부호화 규칙간의 관계를 나타내는 것이다. 표 2에서  는 원천부호화 규칙에 1회 위배되는 문자부호이다. 표 2에서  는 원천부호화 규칙에 2회 위배되는 문자부호들이다. 즉, 표 2에서 보듯이 원천부호규칙에 위배되는 문자부호의 수가 76개이다.

표 2. 사용 빈도율이 높은 한자 문자 150개와 원천부호화 규칙
Table 2. Relation between a higher frequency use Chinese characters code and source coding rule

Cha-racter	code	Cha-racter	code	Cha-racter	code	Cha-racter	code
的	7684	時	56F6	家	5BB6	還	8FD8
一	4E00	就	5C31	种	79CD	定	5B9A
是	662F	出	51FA	里	79CD	實	5B92
不	4E0D	說	8BF4	多	591A	如	5982
了	4E86	會	4F1A	經	7ECF	么	4E48
在	5728	也	4E5F	自	81EA	物	7269
有	6709	子	5B50	現	73B0	法	6CD5
人	4EBA	學	5B66	同	540C	你	4F60
這	8FD9	發	53D1	后	540E	好	597D
上	4E0A	着	7740	産	4EA7	性	6027
大	5927	對	5BF9	方	65B9	民	6C11
來	6765	作	4F5C	工	5DE5	從	4ECE
和	548C	能	80FD	行	884C	天	5929
我	6211	可	53EF	面	9762	化	5316
个	4E2A	于	4E8E	那	90A3	等	7B49
中	4E2D	成	6210	小	5C0F	力	529B
地	5760	用	7528	所	6240	本	672C
爲	4E3A	過	8FC7	起	8D77	長	957F
他	4ED6	動	52A8	去	53BB	心	5FC3
生	751F	主	4E3B	之	4E4B	把	628A
要	8981	下	4E0B	都	90FD	部	90E8
們	4EEC	而	800C	然	7136	義	4E49
以	4EE5	年	5E74	理	7406	樣	6837
到	5230	分	5206	進	8FDB	事	4E8B
國	56FD	得	5F97	体	4F53	看	770B
業	4E1A	沒	6CA1	全	5168	者	8005
当	5F53	想	60F4	制	5236	形	5F62
因	56E0	意	610F	明	660E	應	5E94
高	9AF8	三	4E09	相	76F8	頭	5934
十	5341	只	53EA	兩	4E24	无	65E0
開	5F00	重	91DD	情	60C5	量	91CF
些	4E9B	点	70B9	外	5916	表	8868
社	793E	与	4E0E	間	9614	象	8C71
前	524D	使	4F7F	二	4E8C	气	6C14
又	53C8	但	4F46	關	5173	文	6587
它	5B83	度	5EA6	活	6D3B	展	5C55
水	6C34	由	7531	正	6B63	合	5408
其	5176	道	9063	-	-	-	-

2.3 한자 150개 문자 부호의 사용 빈도율

표 3은 사용빈도율이 높은 상용 한자 문자부호 150개에 대한 사용 빈도율이다. 표 4의 한자 문자부호 150개에 대한 사용빈도율은 전체 사용빈도율의 약 50%를 넘는다.

표 3. 사용 빈도가 높은 한자 150개 문자의 사용 빈도율
Table 3. Frequency rate of use of 105ea Chinese character codes

Cha-racter	frequent rate	Chara cter	frequenct rate	Chara cter	frequent rate	Cha-racter	frequent rate
的	4.8867	時	0.4916	家	0.3241	還	0.2402
一	1.4062	就	0.4905	种	0.3232	定	0.2389
是	1.3155	出	0.4838	里	0.3198	實	0.2374
不	1.0707	說	0.4763	多	0.3178	如	0.2367
了	0.9517	會	0.4695	經	0.3051	么	0.2335
在	0.9258	也	0.4656	自	0.2991	物	0.232
有	0.9082	子	0.4351	現	0.2977	法	0.2316
人	0.7822	學	0.4314	同	0.2958	你	0.2286
這	0.7619	發	0.4224	后	0.2938	好	0.2281
上	0.6031	着	0.3991	産	0.2932	性	0.2272
大	0.5842	對	0.3911	方	0.2843	民	0.2214
來	0.5776	作	0.3842	工	0.2793	從	0.2175
和	0.5765	能	0.3743	行	0.2792	天	0.2166
我	0.576	可	0.3674	面	0.2768	化	0.2159
个	0.5724	于	0.36	那	0.2759	等	0.2151
中	0.546	成	0.3593	小	0.2742	力	0.2126
地	0.5375	用	0.358	所	0.273	本	0.2125
爲	0.5351	過	0.3515	起	0.2651	長	0.2115
他	0.4916	動	0.3512	去	0.2585	心	0.2109
生	0.4905	主	0.3496	之	0.2496	把	0.2091
要	0.4838	下	0.3477	都	0.2485	部	0.2084
們	0.4763	而	0.3433	然	0.2482	義	0.2081
以	0.4695	年	0.3425	理	0.2456	樣	0.2065
到	0.4656	分	0.34	進	0.245	事	0.2041
國	0.4351	得	0.3253	体	0.244	看	0.2037
業	0.1952	沒	0.1783	全	0.16	者	0.149
当	0.1936	想	0.1767	制	0.1579	形	0.1489
因	0.1925	意	0.1736	明	0.1577	應	0.1484
高	0.1892	三	0.1716	相	0.1575	頭	0.1484
十	0.1889	只	0.1708	兩	0.1574	无	0.1477
開	0.1887	重	0.1671	情	0.1564	量	0.1475
些	0.1884	点	0.1662	外	0.156	表	0.1469
社	0.1841	与	0.1627	間	0.1556	象	0.1454
前	0.1828	使	0.1626	二	0.1538	气	0.1436
又	0.1808	但	0.1623	關	0.1536	文	0.1434
它	0.1806	度	0.1615	活	0.1509	展	0.143
水	0.1793	由	0.1603	正	0.1506		
其	0.1791	道	0.1601	合	0.1504		
계	50.1681%						

* 150자의 빈도율은 전체 글자수 20,684자의 사용 빈도율을 100%로 하였을 때의 빈도율이다.

표 4는 표 2와 표 3에서 원천부호화 규칙에 위배되는 문자부호들에 대한 사용 빈도율이다. 번호 1~53까지의 문자부호는 원천부호화 규칙에 한번 위배된 것들이다. 번호 54~77까지는 원천부호화 규칙에 두 번 위배된 문자의 사용 빈도율이다.

표 4. 원천부호화 규칙 위배 문자의 사용 빈도율
Table 4. Frequency rate of a use of violated characters with Source coding rule.

No	Cha-racter	frequent rate	No	Cha-racter	frequent rate	No	Cha-racter	frequent rate
1	的	4.8867	27	民	0.2115	53	气	0.1436
2	了	0.9517	28	長	0.2037	54	文	0.1434
3	這	0.7619	29	樣	0.1968	55	一	1.4062
4	我	0.576	30	業	0.1952	56	不	1.0707
5	地	0.5375	31	当	0.1936	57	有	0.9082
6	生	0.4905	32	因	0.1925	58	上	0.6031
7	要	0.4838	33	高	0.1892	59	對	0.3515
8	到	0.4656	34	十	0.1889	60	能	0.3496
9	國	0.4351	35	社	0.1841	61	成	0.3425
10	就	0.4224	36	它	0.1806	62	下	0.3198
11	說	0.3911	37	水	0.1793	63	而	0.3178
12	也	0.3743	38	其	0.1791	64	分	0.2991
13	子	0.3674	39	三	0.1716	65	得	0.2977
14	着	0.358	40	只	0.1708	66	同	0.2759
15	可	0.3477	41	点	0.1662	67	后	0.2742
16	過	0.3253	42	与	0.1627	68	那	0.2482
17	經	0.2793	43	使	0.1626	69	小	0.2456
18	自	0.2792	44	道	0.1601	70	都	0.2374
19	現	0.2768	45	相	0.1575	71	理	0.2335
20	行	0.2496	46	情	0.1564	72	性	0.2125
21	所	0.245	47	間	0.1556	73	心	0.2021
22	進	0.232	48	合	0.1504	74	部	0.1997
23	還	0.2286	49	形	0.1489	75	看	0.1958
24	如	0.2214	50	无	0.1477	76	開	0.1887
25	你	0.2151	51	量	0.1475	77	者	0.149
26	好	0.2126	52	表	0.1469			
계		28.3298%						

전송효율의 개선에 대한 정량적계산은 표4의 원천부호화 규칙에 위배되는 문자부호들에 대한 사용 빈도 율을 사용하였다. 즉 위배된 문자부호들이 전체 문자 중에서 차지하는 비율을 계산하는 방식으로 하였다.[12] 따라서 전송효율의 개선량은 다음과 같다.

$$\sum_{1}^{54}(\text{사용빈도율}) + \sum_{55}^{77}(\text{사용빈도율}) \times 2$$

$$= (4.8867 + 1.4062 + 1.0707..... + 0.1434 + 0.1504) + (1.4062 + 1.0707... + 0.1187 + 0.149) \times 2$$

$$= 37.25(\%)$$

그러므로 이 원천부호화 규칙에 위배되는 문자부호를 사용빈도가 낮은 문자부호로 대체할 경우에 회선부호화과정에서 기능하는 스크램블링 처리율을 개선시키게 되어 데이터의 전송효율을 제고시키는 효과를 얻게 된다.

III. 결 론

본 논문에서는 Unicode에 규정되어 있는 한중일 통합 한자부호체계와 원천부호화 규칙간의 관계를 연구하였다. 본 연구는 데이터링크 계층의 HDLC 통신규약 내 Flag 비트열과 문자 비트열간의 오인을 방지하는 기능과 AMI/HDB-3 회선부호화와 스크램블링 방식을 적용 할 때에 대하여 연구하였다.

연구를 위한 분석대상 한자 유니코드 문자부호는 사용 빈도율이 전체의 50%를 점유하는 150개 문자부호에 대하여 연구하였다.

연구결과 150개 한자 문자부호 중에서 원천부호화 규칙에 위배되는 문자부호수가 77개로 나타났다. 이 77개의 문자부호에 대한 사용 빈도율은 37.25%였다. 이 원천부호화 규칙에 위배되는 사용빈도가 높은 문자부호를 사용빈도가 상대적으로 낮은 문자부호에 적용할 경우, 사용 빈도 율에 해당하는 정도의 회선부호기 내의 스크램블링 효율을 약 37%정도 개선시킬 수 있는 것으로 나타났다.

본 연구와 관련하여 향후 연구하여야 할 사항은 원천부호화 규칙에 따라 원천부호화 위배 문자부호를 사용빈도가 낮은 문자부호로 대체하는 연구가 필요하다.

참고 문헌

[1] American Standards Association, "American Code (July 6, 1999). for Information Interchange", ASA X3.4- 1963, 17 June, 1963.

[2] 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.

[3] RFC 20 "ASCII format for Network Interchange" October 1969 (http://tools.ietf.org/ html/rfc20)

[4] http://en.wikipedia.org/wiki/EBCDIC

[5] http://en.wikipedia.org/wiki/Unicode

[6] Behrouz A. Forouzan, "Data communications" McGraw Hill Korea, pp. 1031-1032, 2007.

[7] ITU-T Recommendation G.703, "Physical/ electrical characteristics of hierarchical digital interfaces" pp. 24-41, Oct. 1998.

[8] Wan-Pyo Hong, "Coding Rule of Characters by 2 bytes with 4x4 bits to Improve the Transmission Efficiency in Data Communications", The Journal of Korea Navigation Institue, Vol. 15, No. 5, 2011.

[9] http://en.wikibooks.org/wiki/Unicode/Character_reference

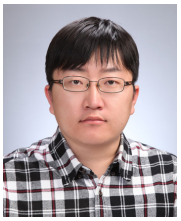
[10] http://www.cncorpus.org. Ministry of Education Institute of applied Linguistice "Modern Chinese corpus work frequency table"

[11] Chang-young Lee, "Improvement of the Linear Predictive Coding with Windowed Autocorrelation", The Journal of The Korea Institute of Electronic Communication Sciences, Vol. 6, No. 2, pp. 186-190, 2011

[12] Wan-Pyo Hong, "Study on the ASCII Code in the side of the Transmission Efficiency in Data Communications", The Journal of Korea Navigation Institute, Vol. 6, No. 5, 2011.

[13] Young-Oh Han, "A study on motion prediction and subband coding of moving pictuers using GRNN", The Journal of The Korea Institute of Electronic Communication Sciences, Vol. 5, No. 3, pp. 255-265, 2010.

저자 소개



태동진(Dong-Zhen Tai)

2002년 한세대학교 컴퓨터공학과 졸업(공학사)
2003년 한세대학교 대학원 IT융합과 공학석사과정
※ 관심분야 : 전파통신, RFID, 문자코딩

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