Regular paper http://dx.doi.org/10.13067/JKIECS.2015.10.7.831

유니코드 기반 UTF-8 한글글자마디 부호의 회선부호기내 스크램블링 발생에 관한 연구

홍완표*

Study on Scrambling Occurrence in Line Coder for UTF-8 Hangul Syllable Code based on Unicode

Wan-Pyo Hong*

요 약

본 논문은 국제적 문자부호체계인 유니코드 체계를 기반으로 한 한글글자마디의 UTF-8부호가 회선부호기상에서 어느 정도 스크램블링이 발생하는 지를 연구하였다. 회선부호기의 ... 경우를 대상으로 하였다. 이것은 ITU-T의 규격중 AMI 회선부호기에 적용되는 HDB-3 스크램블링과 관련된다. 본 논문은 스크램블링이 발생하는 문자의 코드를 분석하기 위해 문자의 원천부호화 규칙을 적용하였다. 연구결과 유니코드를 기반으로 하는 UTF-8한글글자마디부호 중에서 약 39%의 스크램블링이 발생하는 것으로 분석되었다.

ABSTRACT

This paper studied scrambling occurrence in the line coder for UTF-8 hangul syllable code based on unicode. The paper suggested that scrambling is occurred when consecutive four "0" bit is entered into the line coder from the source codes. Currently, ITU-T is applying HDB-3 scrambling method in AMI line coder, According to the study result, scrambling is occurred about 39% in UTF-8 code on Unicoe Hangul syllable.

키워드

UTF-8, Hangul Syllable, Unicode, Source Code, Line Coder, Scrambling
UTF-8, 한글글자마디, 유니코드, 원천부호, 회선부호기, 뒤섞기

Ⅰ. 서 론

정보기기로부터 회선부호기에 입력되는 원천부호는 회선 부호화과정을 거쳐 전송로로 전송된다. 장거리 전송용으로 사용되는 회선부호방식은 AMI(Alternate Mark Inversion)이다[1-2]. 이 방식은 이진수 1비트들이 음과 양의 펄스로 교체되어 전송되고 이진수 0비트들은 무신호(spa ce)형태로 전송된다. 그러나 일정개수 이상의 연속된 0비트의 이진신호가 발생할 경우에는 연속 이진 0신호가 발생하지 않는 부호로 변환하여 전송한다. 이것을 스크램블링이라 한다. ITU-T G.703에 표준화 되어 있는 스크램블링 방식은 HDB-2/3, B6ZS/B8ZS, CMI등이 있다[3-4]. 이 규격에서 HDB-3 방식은 디지틀 인터페이스 E12 (2048 Mbps), E22, E31 (34.368Mbps), B8ZS 방식은 E11 (1.544Mbps), E21 (6.312Mbps) 그리고 CMI방식은 E4 (139.264Mbps), STM-0 (51.840Mbps), STM-1 (111.520Mbps)에서 사용되는 것으로 규격화되어 있다[5].

* 교신저자 (corresponding author) : 한세대학교 정보통신공학과 (wphong@hansei.ac.kr)
접수일자 : 2015. 06. 22 심사(수정)일자 : 2015. 07. 13 게재확정일자 : 2015. 07. 23

대용량 트래픽을 처리하는 광통신 전송망이 확장되면서 CMI(Coded Mark Inversion) 스크램블링 방식이 적용되고 있으나 대용량 광통신망을 사용하지 못하고 있는 상대적으로 저용량 트래픽 전송망이 여전히 존재하고 있는 상황이다.

본 논문은 회선부호기에 이진 0 비트가 연속하여 네 개가 입력되었을 때 스크램블링이 발생하는 HDB-3 스크램블링 방식을 대상으로 하고 있다[6]. 이 원천부호화 규칙은 원천부호에서 발생하는 스크램블링이 발생되지 않는 원천부호를 만들기 위한 것이지만, 본 연구에서는 원천부호내에서 스크램블링이 발생하는 부호수를 산출하는데 사용하였다. 분석된 UTF-8한글글자마디 부호는 총 11,172개의 유니코드 한글글자마디부호를 대상으로 하였다. 유니코드 한글글자마디부호 체계는 16진수 4개로 표현된다(2바이트). 즉 하나의 부호는 16비트를 기준으로 구성되어 있다. UTF-8부호체계는 이 유니코드를 기준으로 하여 부호화된다. 유니코드는 UTF-X부호체계로 변환되기 위해 일정한 규칙에 의하여 포맷된다. 즉 유니코드를 통신망에 전송하기 위해 회선부호기에 입력되기 전에 UTF-8부호체계로 변환시킨다. 한글글자마디의 UTF-8부호체계는 16진수 6개 즉 24비트(3바이트) 체계를 갖고 있다. 즉 유니코드 16비트 3바이트를 기준으로 할 때 8비트 1바이트가 증가된다. 이와 같이 UTF-8부호체계는 유니코드 체계에 의존하고 있지만 별도의 변환 규칙을 가지고 있다. 그러므로 스크램블링 측면에서 볼 때 UTF-8 부호체계는 유니코드와 UTF-8변환규칙에 의해 영향을 받게 된다. 본 연구에서는 전자를 연구대상으로 하였다.

본 연구결과에 의하면 한글글자마디의 경우에 유니코드 자체에서 발생하는 스크램블링에 비하여 UTF-8 부호에서 발생하는 스크램블링이 약간 많은 것으로 나타났다. UTF-8부호체계에서 발생하는 스크램블링은 UTF-8부호체계 뿐만 아니라 유니코드의 부호체계에 의하여도 영향을 받는 것으로 나타났다. 본 연구결과는 향후 스크램블링이 최소화되는 원천부호체계와 UTF-8부호 변환규칙의 개발에 참고가 될 것으로 기대된다.

II. UTF-8부호체계 분석

2.1 문자의 원천부호화 규칙

표 1은 문자의 원천부호화 규칙이다. 왼쪽의 상위비트열을 기준으로 하여 상위비트열에 연결되는 하위비트열의 조합가능여부를 보여 주고 있다. 맨 왼쪽의 16진수는 상위4비트에 대한 것이고 2진수 4비트는 이 16진수 값이다. 하위비트열은 16진수로 나타내고 있다. 조합제한 가능여부는 상위비트열에 연속되는 네 개의 0비트가 만들어지는 하위비트열을 조합제한, 그렇지 않는 비트열을 조합가능으로 나타내고 있다. 예를 들어 상위 16진수가 0인 경우에 이미 자체내에서 0비트가 네 개를 갖고 있으므로 하위 16진수 어느 것이 연결되든 간에 스크램블링이 발생하게 된다. 상위 16진수가 1인 경우에는 하위 16진수 총 16개 중에서 16진수 0의 경우외에는 모두 연결이 가능하다. 이 표 1에서 보듯이 상위16진수가 홀수인 경우에는 하위16진수가 0인 경우를 제외하고는 모두 스크램블링이 발생하지 않는 부호로 조합시킬 수 있다. 상위 16진수 8의 경우가 쌍위 16진수 0을 제외하고 조합이 제한되는 하위 16진수가 가장 많음을 알 수 있다.

표 1. 문자의 원천부호화 규칙[6]

Table 1. Character source coding rule[6]

HEXA	Upper Bits	Lower Bits	
		Composition Limitation	Composition Possible
0	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

2.2 Unicode기반 한글글자마디 부호의 분석

표 2는 유니코드에서 스크램블링 발생하는 부호에 대한 것을 일부 보여 주고 있다[7]. 글자에 밑줄이 있는 것이 스크램블링이 발생하는 글자이다. 이 표 2는 유니코드 총 11,172자를 분석한 것이다. 표 3은 유니코드내 한글글자마디 11,172자내에서 스크램블링이 발생하는 현황이다. 이 표 3은 표 1의 원천부호화 규칙에 의하여 스크램블링이 발생하는 16진수 부호만으로 도출된 것이다. 16진수 두 자리의 왼쪽 숫자는 상위비트열이고 오른쪽 숫자가 하위비트열이다. 왼쪽 숫자를 기준으로 하여 2진수 0비트가 네 개이상 발생하는 부호들이다. 이 표에서 16진수 00부호가 89개 있음을 나타낸다. 이 스크램블링 발생수는 한 개의 문자부호인 4개의 16진수 간에 스크램블링이 발생하는 16진수가 겹치는 것을 포함하는 개수이다. 예를 들어 B006부호의 경우에 B0와 06에 의해 두 번 스크램블링이 발생하는데 이 표의 경우에는 B0, 00, 06과 같이 도출된 것이다. 이와 같이 도출된 유니코드 한글글자마디 11,172개에서 발생하는 스크램블링 횟수는 총 5,657번이 되는 것으로 나타났다. 글자마디 한 개의 부호를 기준으로 하였을 때 총 4,015개의 부호에서 스크램블링이 발생하였다. 즉 스크램블링이 두 개의 숫자간에 겹쳐서 발생하는 부호수가 1,642개가 된다. 스크램블링이 발생하는 총4,015개의 부호중에서 한 개의 부호내에서 스크램블링이 연속하여 두 번 발생하는 글자마디는 총291개로 나타났다. 이중에서 B000, C000, D000부호의 경우에는 스크램블링이 세 번 발생한다. 즉 글자마디를 기준으로 할 때 한글글자마디 전체를 회선부호기에 입력시킨다고 가정할 경우에 총 4,309번의 스크램블링이 발생하게 된다. 이것은 전체 글자마디수와 비교할 때 의 약38.6%에 해당된다.

표 2. 유니코드기반 한글글자마디 부호의 스크램블링 발생 예

Table 2. Examples of scrambling occurrence of hangul syllable code based on unicode

Code	Han-gul	Code	Han-gul	Code	Han-gul
B006	꼦	B406	됆	B706	뛎
B007	꼧	B407	됇	B707	뛏
B008	꼨	B408	됈	B708	뛐
B009	꼩	B409	됉	B709	뛑
B00A	꼪	B40A	됊	B70A	뛒
B00B	꼫	B40B	됋	B70B	뛓
B00C	뀌	B40C	됌	B70C	뜌
B00D	뀍	B40D	됍	B70D	뜍
B00E	뀎	B40E	됎	B70E	뜎
B00F	뀏	B40F	됏	B70F	뜏
B010	뀐	B410	됐	B710	뜐
B011	뀑	B411	됑	B711	뜑
B012	뀒	B412	됒	B712	뜒
B013	뀓	B413	됓	B713	뜓
B014	뀔	B414	됔	B714	뜔
B015	뀕	B415	됕	B715	뜕

*____ : Character and Code occuring Scrambling

유니코드 한글글자마디에서 스크램블링이 발생하는 문자수가 4,015개이므로 스크램블링이 발생하지 않는 글자마디는 총7,157개가 된다. 2005년도에 한국국립국어원에서 연구 발표한 한국어 글자마디 사용현황에 의하면 한국어어휘를 구성하고 있는 글자마디가 총 1,540자이다[7-10]. 이것은 유니코드에서 스크램블링이 발생하지 않는 코드에 이 글자 모두를 부호화하여도 총 5,617개의 스크램블링이 발생하지 않는 부호점을 갖게 됨을 의미한다[11]. 즉 유니코드 한글글자마

디 부호체계를 이러한 부호체계로 변환하여 회선부호기에 입력시킨다면 스크램블링이 전혀 발생되지 않게 됨을 의미하는 것이다.

표 3. 유니코드 한글글자마디 부호의 스크램블링 발생 현황

Table 3. Situation of scrambling occurrence in unicode hangul syllable

Code (Hexa)	Number of Scrambling	Code (Hexa)	Number of Scrambling
00	89	50	92
01	92	60	92
02	92	61	92
03	92	70	92
04	92	80	76
05	92	81	76
06	92	82	76
07	92	83	76
08	92	84	76
09	92	85	76
0A	92	86	76
0B	92	87	76
0C	92	90	76
0D	92	A0	76
0E	92	A1	76
0F	92	B0	331
10	92	C0	347
20	92	C1	347
21	92	C2	347
30	92	C3	347
40	92	D0	347
41	92	E0	91
42	92	E1	91
43	92	F0	91
Total	2205	Total	3543
Total 5657			

2.3 Unicode기반 UTF-8 한글글자마디부호 분석

표 4는 유니코드를 UTF-8부호체계로 변환한 11,172개의 한글글자마디의 일부를 보여 주는 것이다. 글자마디에 밑즐이 있는 글자의 코드가 스크램블링이 발생하는 코드이다. 표 5는 표 1의 원천부호화 규칙을 사용하여 작성된 표 4으로부터 추출된 것으로 UTF-8 부호체계내에서 스크램블링이 발생하는 횟수를 분석한 것이다. 이 표 5내의 괄호는 스크램블링은 발생하나 스크램블링이 발생하는 다른 값과 겹치는 것들이다. 예를 들어 한글글자마디 각의 부호 eab081의 경우처럼 "08"은 "b0"와 겹치게 된다. 그러므로 표1의 원천부호화 규칙에 의한 스크램블링 발생상태는 괄호내의 숫자를 포함한 총 4,949번이 된다. 여기서 실제로 스크램블링이 발생하는 횟수는 이 표에서와 같이 UTF-8 한글글자마디 총11,172개의 부호내에서 4,245회 나타났다. 이것을 총한글글자마디와 비교하면 약 39%수준이 된다. 한편 한글글자마디부호중에서 스크램블링이 발생하는 글자마디는 총 3,622자로서 전체의 32.4%에 달한다. 이것은 스크램블링이 발생하는 총 3,622자중에서 623개의 글자마디부호에서는 스크램블링이 연속하여 두 번 발생하는 것을 나타내는 것이다.

표 4. 유니코드기반 UTF-8 한글글자 마디 부호의 스크램블링 발생 분석

Table 4. An analysis on scrambling occurrence of UTF-8 hangul syllables code based on unicode

Hangul Syllable	UTF-8 code	Hangul Syllable	UTF-8 code	Hangul Syllable	UTF-8 code	Hangul Syllable	UTF-8 code
가	eab080	관	eab480	글	eab880	꼀	eabc80
각	eab081	괁	eab481	긁	eab881	꼁	eabc81
갂	eab082	괂	eab482	긂	eab882	꼂	eabc82

갂	eab083	괃	eab483	긊	eab883	꼃	eabc83
간	eab084	괄	eab484	긋	eab884	껌	eabc84
갅	eab085	괅	eab485	긎	eab885	껍	eabc85
갆	eab086	괆	eab486	긒	eab886	껎	eabc86
갇	eab087	괇	eab487	긓	eab887	껏	eabc87
갈	eab088	괈	eab488	금	eab888	껐	eabc88
갉	eab089	괉	eab489	급	eab889	껑	eabc89
갊	eab08a	괊	eab48a	긚	eab88a	껒	eabc8a
~							
폀	ed8f80	퓀	ed9380	햀	ed9780	횤	ed9b80
폁	ed8f81	퓁	ed9381	행	ed9781	횥	ed9b81
폂	ed8f82	퓂	ed9382	햆	ed9782	횦	ed9b82
폃	ed8f83	퓃	ed9383	햇	ed9783	횧	ed9b83
폄	ed8f84	퓄	ed9384	했	ed9784	후	ed9b84
폅	ed8f85	퓅	ed9385	헀	ed9785	훅	ed9b85
폆	ed8f86	퓆	ed9386	헂	ed9786	훆	ed9b86
폇	ed8f87	퓇	ed9387	행	ed9787	훇	ed9b87
폈	ed8f88	퓈	ed9388	허	ed9788	훈	ed9b88
평	ed8f89	퓉	ed9389	헉	ed9789	훉	ed9b89
폊	ed8f8a	퓊	ed938a	헊	ed978a	훊	ed9b8a

*_____ : Character and Code occuring Scrambling

표 5. 유니코드기반 UTF-8 한글글자마디 부호의 스크램블링 발생 횟수

Table 5. The number of scrambling occurrence of UTF-8 hangul syllable code based on unicode

Code (Hexa)	Number of Scrambling	Code (Hexa)	Number of Scrambling
00	-	50	-
01	-	60	-
02	-	61	-
03	-	70	-
04	-	80	365
05	-	81	364
06	-	82	364
07	-	83	364
08	(176)	84	364
09	(176)	85	364
0A	(176)	86	364
0B	(176)	87	364
0C	-	90	367
0D	-	A0	301
0E	-	A1	301
0F	-	B0	363
10	-	C0	-
20	-	C1	-
30	-	C2	-
40	-	C3	-
41	-	D0	-
42	-	E0	-
43	-	E1	-
		F0	-
	(704)		4245
Total	4,949		

III. 결 론

본 논문에서는 유니코드의 한글글자마디와 이 부호에 대한 UTF-8부호에서 스크램블링이 발생하는 것에 대하여 연구하였다. 연구결과 유니코드와 UTF-8 한글글자마디 부호를 회선부호기에 입력되는 부호라고 했을 때 각각의 부호체계에서 발생하는 스크램블링 현황은 다음과 같다.

유니코드 부호체계에서 발생하는 스크램블링 횟수는 총 4,306번으로 나타났다. 스크램블링 발생하는 한글글자마디부호수로는 총 4,015개로서 한 부호당 두 번의 스크램블링이 발생하는 부호가 291개이고 세 번 발생하는 부호가 3개로 나타났다. 결과적으로 스크램블링이 발생하는 율은 발생횟수로는 약38.6%, 부호수로는 약36%에 달했다. UTF-8 부호체계에서는 총4,245번의 스크램블링이 발생하고 부호수로는 총 3,622개에서 스크램블링이 발생했다. 즉 한 개의 부호에서 두 번 스크램블링이 발생하는 부호수는 623개였다. 즉 UTF-8부호체계에서 스크램블링이 발생하는 율은 각각 39%와 32.4%로 나타났다.

본 연구결과는 향후 스크램블링이 최소화 또는 발생되지 않는 한글글자마디 부호체계를 발전시켜서 회선부호기의 데이터 처리품질을 제고시키는데 기여할 것으로 판단된다.

Reference

[1] B. A. Forouzan, *Data communications 5E.* New York: McGraw Hill, 2013

[2] ITU-T, "Physical/Electrical characteristic of hierarchical digital interfaces," *ITU-T G.701,* Geneva, Switzerland, 2001, p. 16.

[3] Behrouz A. Forouzan, "Data communications 5E," McGraw Hill NY, 2013

[4] ITU-T, "Physical/Electrical characteristic of hierarchical digital interfaces," *ITU-T G.701,* Geneva, Switzerland, 2001, p. 48.

[5] ITU-T, "Physical/Electrical characteristic of hierarchical digital interfaces," *ITU-T G.701,* Geneva, Switzerland, 2001, pp. 13-42.

[6] W. Hong, "Coding Rule of Characters by 2 bytes with 4x4 bits to Improve the Transmission Efficiency in Data Communications," *J. of Korea Navigation Institute,* vol. 15, no. 5, 2011, pp. 749-756.

[7] W. Hong, "A Study on the Hangul Syllables of Unicode System considering Data Transmission Efficiency," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 10, no. 1, 2015, pp. 39-46.

[8] W. Hong, "Analysis of Korean Language to optimize the Hangul Character Coding for Information processing and Communication" *J. of the Korea Institute of Electronic Communication Sciences,* vol. 10, no. 3, 2015, pp. 375-380.

[9] Y. Han, "A study on motion prediction and subband coding of moving pictuers using GRNN," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 5, no. 3, 2010, pp. 256-261.

[10] K. Lee, and Y. Son, "Fast Encoding Algorithm of Low Density Codes," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 9, no. 4, 2014, pp. 403-408.

[11] Y. Kim, "A Study on Fractal Image Coding," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 7, no. 3, 2012, pp. 559-566.

저자 소개

홍완표(Wan-Pyo Hong)

1991년 서울과학기술대학교 전자공학과(공학사)
1994년 연세대학교 공학대학원 전자공학전공(공학석사)
1999년 광운대학교 대학원 전자공학과(공학박사)
1990년 전기통신기술사합격
1991년 정보통신부 5급특별채용고시합격 본부 통신정책실, 전파방송관리국, 정보화기획실
1997년 삼성전자(주) 통신사업부 전송영업그룹장
1999년 광운대학교 연구전담교수
2000년 한국정보통신기술사협회장
2002년 한세대학교 정보통신공학과 교수
2014년 USC 동북아언어문화학과 방문학자
※관심분야 : 위성통신방송/문자코딩/통신정책/