Unicode의 UTF-8 부호화 방식의 HDB-3 스크램블링 방식과의 적합성

홍완표*

Compatibility of UTF-8 Encoding System to HDB-3 Scrambling Method

Wan-Pyo Hong*

요 약

본 논문에서는 국내 표준규격인 HDB-3 스크램블링 방식을 기준으로 유니코드(Unicode)의 한글낱자 및 호환용 한글낱자 부호와 이 부호가 UTF-8 부호체계로 변환되었을 때 원천부호화 규칙에 얼마나 부합되는지 여부를 분석하였다. 연구결과 유니코드 한글낱자 및 호환용 한글 낱자 부호체계와 UTF-8부호체계내에 문자의 원천 부호화 규칙에 위배되는 부호가 상당히 존재하는 것으로 나타났다. 특히 UTF-8로 변환함에 따라 그 위배율이 증가하는 것으로 분석되었다.

ABSTRACT

This paper studied how much influence UTF coded data is given to HDB-3 operation efficiency. This paper applied the Hangul Jamo and Compatability Hangul Jamo in Unicode to compare to its UTF-8 coded data. As a result of study, When Unicode is reformatted to UTF-8 code, the data code run counter to the source coding rule was very much increased.

키워드

원천부호화, 회선부호화, HDB-3, UTF, 유니코드

Source coding, Line coding, HDB-3, UTF, Unicode

I. 서 론

데이터통신에 있어서 전송되는 데이터비트열에 일정개수 이상의 “0”의 비트가 연속하여 발생하는 것은 바람직하지 않다. 일정개수 이상의 연속 “0”의 비트가 발생할 경우 동기를 잃을 수 있기 때문이다[1]. 특히 OSI계층의 물리계층에서 회선부호화를 위해 AMII(Alternate Mark Inversion)방식을 사용할 경우에는 더 그러하다. 그러므로 AMII회선부호화방식에서는 연속 “0”의 비트가 발생되는 것을 방지하기 위해 스크램블링 방법을 적용하고 있다. 스크램블링 방식으로는 B8ZS 방식[2]과 HDB-3 방식[3]이 대표적인 방법이다. 전자는 미국표준방식으로 연속“0”의 비트가 8개 발생할 때 이 비트열을 사전에 정해진 일정한 비트열로 대체하는 방식이다[4]. 후자는 ITU-T[5]와 국내 표준방식[6]으로 연속 “0”의 비트가 4개 발생할 때 이 비트열을 사전에 정해진 비트열로 대체하는 방식이다. 그러므로 스크램블링이 발생하는 횟수는 OSI 표현계

* 한세대학교 정보통신공학과(wphong@hansei.ac.kr)
접수일자 : 2012. 11. 25 심사(수정)일자 : 2012. 12. 30 게재확정일자 : 2013. 02. 20

표 1. 유니코드 한글낱자 부호와 원천부호화 규칙
Table 1. Unicode Hangul Jamo and source coding rule

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
1100	ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ	ㅁ	ㅂ	ㅃ	ㅅ	ㅆ	ㅇ	ㅈ	ㅉ	ㅊ	ㅋ
1110	ㅌ	ㅍ	ㅎ	ㅏ	ㅐ	ㅑ	ㅒ	ㅓ	ㅔ	ㅕ	ㅖ	ㅗ	ㅘ	ㅙ	ㅚ	ㅛ
1120	ㅜ	ㅝ	ㅞ	ㅟ	ㅠ	ㅡ	ㅢ	ㅣ	ㆍ	ᆞ	ᆢ	ᄀᆞ	ᄁᆞ	ᄂᆞ	ᄃᆞ	ᄅᆞ
1130	ᄆᆞ	ᄇᆞ	ᄈᆞ	ᄉᆞ	ᄊᆞ	ᄋᆞ	ᄌᆞ	ᄍᆞ	ᄎᆞ	ᄏᆞ	ᄐᆞ	ᄑᆞ	ᄒᆞ	ᅀᆞ	ᅁᆞ	ᅂᆞ
1140	ᅃᆞ	ᅄᆞ	ᅅᆞ	ᅆᆞ	ᅇᆞ	ᅈᆞ	ᅉᆞ	ᅊᆞ	ᅋᆞ	ᅌᆞ	ᅍᆞ	ᅎᆞ	ᅏᆞ	ᅐᆞ	ᅑᆞ	ᅒᆞ
1150	ᅓᆞ	ᅔᆞ	ᅕᆞ	ᅖᆞ	ᅗᆞ	ᅘᆞ	ᅙᆞ	ᅚᆞ	ᅛᆞ	ᅜᆞ	ᅝᆞ	ᅞᆞ				
1160		ᅡ	ᅢ	ᅣ	ᅤ	ᅥ	ᅦ	ᅧ	ᅨ	ᅩ	ᅪ	ᅫ	ᅬ	ᅭ	ᅮ	ᅯ
1170	ᅰ	ᅱ	ᅲ	ᅳ	ᅴ	ᅵ	ᅶ	ᅷ	ᅸ	ᅹ	ᅺ	ᅻ	ᅼ	ᅽ	ᅾ	ᅿ
1180	ᆀ	ᆁ	ᆂ	ᆃ	ᆄ	ᆅ	ᆆ	ᆇ	ᆈ	ᆉ	ᆊ	ᆋ	ᆌ	ᆍ	ᆎ	ᆏ
1190	ᆐ	ᆑ	ᆒ	ᆓ	ᆔ	ᆕ	ᆖ	ᆗ	ᆘ	ᆙ	ᆚ	ᆛ	ᆜ	ᆝ	ᆞ	ᆟ
11A0	ᆠ	ᆡ	ᆢ						ᆨ	ᆩ	ᆪ	ᆫ	ᆬ	ᆭ	ᆮ	ᆯ
11B0	ᆰ	ᆱ	ᆲ	ᆳ	ᆴ	ᆵ	ᆶ	ᆷ	ᆸ	ᆹ	ᆺ	ᆻ	ᆼ	ᆽ	ᆾ	ᆿ
11C0	ᇀ	ᇁ	ᇂ		ᇃ	ᇄ	ᇅ	ᇆ	ᇇ	ᇈ	ᇉ	ᇊ	ᇋ	ᇌ	ᇍ	ᇎ
11D0	ᇏ	ᇐ	ᇑ	ᇒ	ᇓ	ᇔ	ᇕ	ᇖ	ᇗ	ᇘ	ᇙ	ᇚ	ᇛ	ᇜ	ᇝ	ᇞ
11E0	ᇟ	ᇠ	ᇡ	ᇢ	ᇣ	ᇤ	ᇥ	ᇦ	ᇧ	ᇨ	ᇩ	ᇪ	ᇫ	ᇬ	ᇭ	ᇮ
11F0	ᇯ	ᇰ	ᇱ	ᇲ	ᇳ	ᇴ	ᇵ	ᇶ	ᇷ	ᇸ						

참고문헌[9]표5와 6중복위배 부호 참고문헌[9]표5위배 부호 참고문헌[9]표6위배부호

층에서 정보기기에 입력되는 문자나 기호를 어떻게 부호화하느냐에 따라 영향을 주게 된다[7]. 즉 표현계층에서 입력되는 정보를 원천 부호화할 때 각 문자나 기호의 부호를 구성하는 각 문자나 기호의 부호 및 이 각 부호가 상호 연결될 때 연속“0”의 비트가 발생되지 않도록 할 필요가 있다. 참고문헌 [8][9]에서는 이점을 고려한 원천 부호화 규칙을 연구하여 제시하였다. 데이터통신시스템은 유니코드(Unicode)를 전송할 때 UTF(Unicode Transformation Format) 부호체계로 변환하여 전송한다[11]. 본 논문에서는 유니코드가 UTF 부호화될 때 연속 “0”의 비트열을 발생하는 것과 어떠한 상관관계가 있는지를 연구하였다. 본 논문에서는 한글낱자와 한글글자마디 유니코드를 연구대상으로 하였다. 본 연구에서는 최근 인터넷에서 그 사용이 증가[12]하고 있는 UTF-8 부호[13]에 대하여 연구하였다. UTF-8은 유니코드에 대한 가변길이 문자 부호화 체계이다.

II. 유니코드의 한글낱자와 원천부호화

2.1 한글낱자 유니코드와 원천 부호화규칙

표 1과 표 2는 유니코드 BMP에 있는 한글낱자(Hangul Jamo)[14]와 호환용 한글낱자[15]에 대한 부호표이다. 표 1의 한글낱자 유니코드는 1993년 6월에 유니코드 버전 1.1(ISO/IEC 10646-1:1993)로 추가되었다[16]. 표 2의 호환용 한글 낱자는 1991년 10월에 유니코드 버전 1.0에 추가되었다[17]. 표 1에서 U+1100-U+115E까지는 초성, U+1161-11A7까지는 중성, U+11A8-U+11FF까지는 종성에 대한 부호이다. 이 부호표는 한글 옛 체를 포함하고 있고 초성, 중성, 종성을 ㄱ, ㄴ,ㅏ,ㅑ 의 자음과 모음 순으로 부호를 부여하고 있다.

표 2는 유니코드상의 호환용 한글 글자에 대한 부호표이다. 이 부호표는 자음과 모음에 대한 것으로 현재는 사용하지 않는 옛글을 포함하고 있다. 표 1과 표 2에서 보듯이 자음과 모음의 배열은 자음과 모음의 순서에 따라 단순 배열되어 있음을 알 수 있다.

본 논문에서는 참고문헌 [9]에서 제시한 문자의 원천부호화에 대한 규칙을 적용하여 유니코드 한글낱자와 호환용 한글낱자에 대하여 분석하였다.

표 2. 유니코드 호환용 한글낱자 부호와 원천부호화 규칙
Table 2. Hangul compatibility Jamo and source coding rule

U+	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
3130		ㄱ	ㄲ	ㄳ	ㄴ	ㄵ	ㄶ	ㄷ	ㄸ	ㄹ	ㄺ	ㄻ	ㄼ	ㄽ	ㄾ	ㄿ
3140	ㅀ	ㅁ	ㅂ	ㅃ	ㅄ	ㅅ	ㅆ	ㅇ	ㅈ	ㅉ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ	ㅏ
3150	ㅐ	ㅑ	ㅒ	ㅓ	ㅔ	ㅕ	ㅖ	ㅗ	ㅘ	ㅙ	ㅚ	ㅛ	ㅜ	ㅝ	ㅞ	ㅟ
3160	ㅠ	ㅡ	ㅢ	ㅣ		ㅥ	ㅦ	ㅧ	ㅨ	ㅩ	ㅪ	ㅫ	ㅬ	ㅭ	ㅮ	ㅯ
3170	ㅰ	ㅱ	ㅲ	ㅳ	ㅴ	ㅵ	ㅶ	ㅷ	ㅸ	ㅹ	ㅺ	ㅻ	ㅼ	ㅽ	ㅾ	ㅿ
3180	ㆀ	ㆁ	ㆂ	ㆃ	ㆄ	ㆅ	ㆆ	ㆇ	ㆈ	ㆉ	ㆊ	ㆋ	ㆌ	ㆍ	ㆎ	

참고문헌[9]표5와 6중복위배 부호 참고문헌[9]표5위배 부호 참고문헌[9]표6위배부호

2.2 유니코드 한글낱자 부호와 원천부호화 규칙

표 1은 유니코드의 한글낱자 부호와 참고문헌[9]에서 제시하고 있는 원천부호화 규칙간의 관계를 보여주는 것이다. 이 부호표는 한글의 자음과 모음 낱자를 초성, 중성, 종성으로 구분하여 부호화하고 있다. 이 표 1에서 보듯이 이 부호표의 BMP 부호판은 1100~11FF이다. 즉 전체 부호수는 256개가 된다. 결과적으로 총 256개의 부호수에서 63개의 부호가 참고문헌[9] 표 5와 표 6의 원천부호화 규칙에 위배됨을 알 수 있

표 3. 한글낱자(자모) 사용빈도수, 빈도율 및 총 빈도율
Table 3. Hangul Jamo using rate

초성	hexa	빈도수	빈도율(%)	총빈도율(%)	중성	hexa	빈도수	빈도율(%)	총빈도율(%)	종성	hexa	빈도수	빈도율(%)	총빈도율(%)
ㄷ	1100	308480	23.45	10.27	ㅏ	1161	535428	41.54	17.86	ㄴ	11AB	88733	22.32	2.98
ㅇ	110B	232752	17.69	7.75	ㅣ	1175	187794	14.57	6.26	ㄹ	11AF	72026	18.12	2.49
ㄱ	1100	175213	13.32	5.83	ㅓ	1165	132439	10.27	4.42	ㅇ	11BC	66552	16.74	2.30
ㅅ	1109	101877	7.74	3.39	ㅗ	1169	122451	9.50	4.08	ㅆ	11BB	36755	9.25	1.27
ㅈ	110C	98303	7.47	3.27	ㅜ	116E	99602	7.73	3.32	ㄱ	11A8	31752	7.99	1.10
ㅎ	1112	89829	6.83	2.99	ㅡ	1173	98447	7.64	3.28	ㅁ	11B7	29452	7.41	1.02
ㄴ	1102	72781	5.53	2.42	ㅐ	1162	45011	3.49	1.50	ㄴㅎ	11AD	13040	3.28	0.45
ㅁ	1106	64632	4.91	2.15	ㅕ	1167	33685	2.61	1.12	ㅎ	11C2	12992	3.27	0.45
ㅂ	1107	57483	4.37	1.91	ㅔ	1166	16696	1.30	0.56	ㅂㅅ	11B9	10072	2.53	0.35
ㄹ	1105	49339	3.75	1.64	ㅖ	1168	7802	0.61	0.26	ㅂ	11B8	8862	2.23	0.31
ㄸ	1104	23752	1.81	0.79	ㅛ	116D	6866	0.53	0.23	ㅌ	11C0	7785	1.96	0.27
ㅊ	110E	12917	0.98	0.43	ㅠ	1172	1764	0.14	0.06	ㅍ	11C1	4532	1.14	0.16
ㅍ	1111	7728	0.59	0.26	ㅑ	1163	1068	0.08	0.04	ㄷ	11AE	4159	1.05	0.14
ㅆ	110A	6720	0.51	0.22						ㅅ	11BA	3559	0.90	0.12
ㅌ	1110	5132	0.39	0.17						ㅈ	11BD	3268	0.82	0.11
ㅋ	110F	3862	0.29	0.13						ㄴㅈ	11AC	1029	0.26	0.04
ㄲ	1101	2757	0.21	0.09						ㅊ	11BE	1024	0.26	0.04
ㅃ	1108	2019	0.15	0.07						ㄹㅁ	11B1	1013	0.25	0.04
										ㄹㄱ	11B0	2098	0.23	0.03
										ㅋ	11BF	130	0.03	0.00
계		1315576		43.80	계		1289053		42.92	계		398833		13.28
총빈도수 : 3,003,462														

다. 즉 원천 부호화 할 때 조합이 가능한 부호는 총 193개가 된다. 원천부호화 규칙에 위배된다는 것은 2바이트 원천부호에 연속 4개 이상의 “0”의 비트열이 1개 이상 있다는 것을 의미한다. 예를 원천부호화 규칙에 위배되는 것으로 나타난 “ㄱ” 과 “ㄴ”의 경우에 부호가 각각 2진수 0001000100000000 및 0001000100000010로 구성됨을 알 수 있다.·

2.3 유니코드 호환용 한글낱자 부호와 원천부호화 규칙

표 2는 유니코드 호환용 한글낱자에 대한 부호표이다. 이 부호표의 BMP 부호판은 3130~318F까지이다. 이 호환용 한글낱자 부호표는 초성, 중성과 종성으로 구분하지 않고 자음과 모음으로 구분하여 부호화하였다. 이 부호표의 전체 부호수는 96개이다. 이중에서 참고문헌[9]의 표 5와 표6의 원천부호화 규칙에 부합되는 부호는 총 68개가 된다.

2.4 한글 단어 빈도수와 빈도율

표 3은 표 1에 있는 총 254개 낱자중에서 48개에 대해서만 분석한 것이다. 이 48개는 국립국어원에서 조사한 현대국어사용빈도조사결과[18]에 있는 총58,437개의 단어 중에서 사용빈도수가 1000번 이상 되는 192개의 단어에 사용된 낱자들이다. 즉 사용빈도가 1000번 이하의 단어의 낱자에 대하여는 분석하지 않았다. 58,437개 단어의 총 사용빈도수는 1,484,463번이다[18].

빈도수 1000개이상인 192개 단어의 총 빈도수는 519,363개이다. 즉 빈도수 1,000개이상인 192개의 단어가 전체 58,437개 단어에서 차지하는 비율은 0.3%이지만 사용 빈도율은 전체의 35%에 달한다. 이 표 3에서 빈도율은 초성, 중성 및 종성의 각 성내에서의 상대비율이다. 총 빈도율은 총 낱자 빈도수 3,003,462를 기준으로 한 상대비율이다.

그림 1은 사용단어 전체에 대한 사용분야별 빈도수를 나타내는 것이다. 교양분야의 사용빈도수가 가장 많고 다음 신문, 문학 및 잡지 순이다.

그림 2는 단어의 사용빈도수가 1000번 이상되는 단어에 대한 분야별 사용빈도수를 보여 주는 것이다. 사용빈도수가 1000번이상되는 단어의 수는 총 192개로서 이에 대한 총 사용빈도수는 519,363번이다. 분야별 사용빈도를 보면 그림 1과는 약간 다르게 교양, 문학, 신문, 잡지의 순임을 알 수 있다.

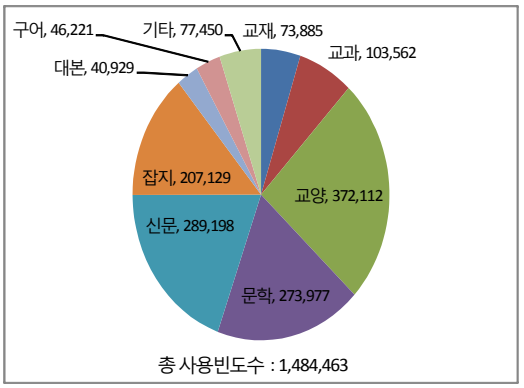

그림 1. 한글 단어 전체 사용빈도수
Fig. 1 Hangul word using number

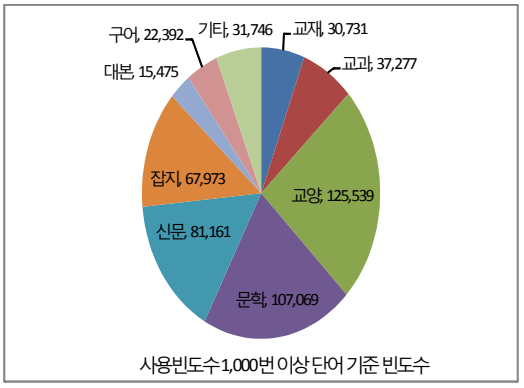

그림 2. 1,000번 이상 사용하는 단어 기준 사용 빈도수
Fig. 2 Hangul word using number with 192 words over 1,000 using number

그림 3은 각 분야별 전체 단어 사용빈도수 대 사용빈도수가 1000개 이상되는 단어의 분야별 사용빈도수에 대한 비이다. 그림에서 보는 바와 같이 사용빈도수가 1000개를 넘는 단어의 사용빈도수가 전체 단어 사용빈도수의 35에 해당됨을 알 수 있다. 분야별로는 구어분야 48, 교재분야 42, 교과분야 36, 문학분야 39 등의 순이다.

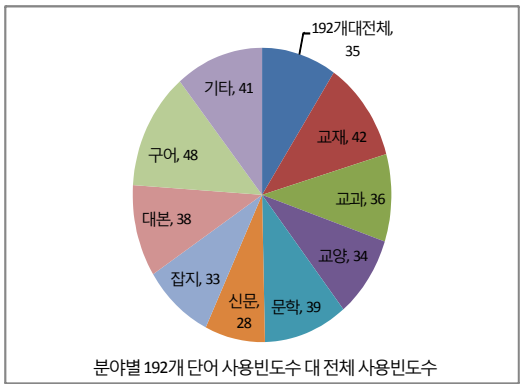

그림 3. 192개 단어 사용빈도수 대 한글 전체 사용빈도수
Fig. 3 192 words using number to Hangul total using number

2.5 유니코드 한글낱자 48개와 원천부호화 규칙

표 4는 한글낱자 48개의 부호와 원천부호화 규칙에 대하여 보여주는 것이다. 이 표에서 보듯이 한글낱자 48개중에서 참고문헌[9]의 표5의 원천 부호화 규칙에 위배되는 부호는 총21개에 해당된다. 즉 48개중 41.2에 해당되는 부호가 원천부호화 규칙에 위배된다. 이것을 초성, 중성 및 종성으로 구분하여 보면 초성에서 원천부호화 규칙에 위배되는 개수는 ㄱ에 두 개를 포함하여 ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ ㅃ ㅅ ㅇ ㅈ ㅊ ㅋ ㅌ에 각 한 개씩 총 17개중에 16개, 중성은 14개중에"ㅏ"에 한 개, 종성은 총 20개 중에서 ㅌㅍㅎ 및 "ㄹㄱ"에 각 한 개 씩 4개가 있다. 이 원천부호화 규칙 위배율을 표 3의 사용 빈도율을 적용하여 계산하면 초성은 94, 중성은 42.12 그리고 종성은 2에 해당된다. 이것을 참고문헌[9]의 표6의 바이트간 원천부호화 규칙에 위배되는 15개의 부호를 포함한다면 48개중 36개가 원천 부호화 규칙에 위배된다. 이것은 전체의 70.6에 해당된다. 표 1의 유니코드 한글낱자 전체 부호수는 256개이다. 이중에서 원천부호화 규칙에 부합되는 부호는 총117개 있다. 이중에서 표 4에 있는 원천부호화 규칙에 부합되는 부호가 17개이다. 그러므로 총100개의 원천부호화에 부합되는 부호가 있는 것이다. 이것은 표 4의 원천부호화 규칙에 위배되는 21개의 부호를 대체하고서도 79개의 가용부호가 남는 것이다.

표 4. 유니코드 한글낱자 48개 부호와 원천부호화 규칙
Table 4. Unicode Hangul Jamo 48ea and source coding rule

문자	hexa	유니코드 부호표															
ㄱ	1100	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
ㄲ	1101	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
ㄴ	1102	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0
ㄷ	1100	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	1
ㄸ	1104	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0
ㄹ	1105	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	1
ㅁ	1106	0	0	0	1	0	0	0	1	0	0	0	0	0	1	1	0
ㅂ	1107	0	0	0	1	0	0	0	1	0	0	0	0	0	1	1	1
ㅃ	1108	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0
ㅅ	1109	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	1
ㅆ	110A	0	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0
ㅇ	110B	0	0	0	1	0	0	0	1	0	0	0	0	1	0	1	1
ㅈ	110C	0	0	0	1	0	0	0	1	0	0	0	0	1	1	0	0
ㅊ	110E	0	0	0	1	0	0	0	1	0	0	0	0	1	1	1	0
ㅋ	110F	0	0	0	1	0	0	0	1	0	0	0	0	1	1	1	1
ㅌ	1110	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0
ㅍ	1111	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
ㅎ	1112	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
ㅏ	1161	0	0	0	1	0	0	0	1	0	1	1	0	0	0	0	1
ㅐ	1162	0	0	0	1	0	0	0	1	0	1	1	0	0	0	1	0
ㅑ	1163	0	0	0	1	0	0	0	1	0	1	1	0	0	0	1	1

ㅋ	1165	0	0	0	1	0	0	0	1	0	1	1	0	0	1	0	1
ㅌ	1166	0	0	0	1	0	0	0	1	0	1	1	0	0	1	1	0
ㅍ	1167	0	0	0	1	0	0	0	1	0	1	1	0	0	1	1	1
ㅎ	1168	0	0	0	1	0	0	0	1	0	1	1	0	1	0	0	0
ㅗ	1169	0	0	0	1	0	0	0	1	0	1	1	0	1	0	0	1
ㅛ	116D	0	0	0	1	0	0	0	1	0	1	1	0	1	1	0	1
ㅜ	116E	0	0	0	1	0	0	0	1	0	1	1	0	1	1	1	0
ㅟ	1172	0	0	0	1	0	0	0	1	0	1	1	1	0	0	1	0
ㅡ	1173	0	0	0	1	0	0	0	1	0	1	1	1	0	0	1	1
ㅣ	1175	0	0	0	1	0	0	0	1	0	1	1	1	0	1	0	1
ㄱ	11A8	0	0	0	1	0	0	0	1	1	0	1	0	1	0	0	0
ㄴ	11AB	0	0	0	1	0	0	0	1	1	0	1	0	1	0	1	1
ㄵ	11AC	0	0	0	1	0	0	0	1	1	0	1	0	1	1	0	0
ㄶ	11AD	0	0	0	1	0	0	0	1	1	0	1	0	1	1	0	1
ㄷ	11AE	0	0	0	1	0	0	0	1	1	0	1	0	1	1	1	0
ㄹ	11AF	0	0	0	1	0	0	0	1	1	0	1	0	1	1	1	1
ㄺ	11B0	0	0	0	1	0	0	0	1	1	0	1	1	0	0	0	0
ㄻ	11B1	0	0	0	1	0	0	0	1	1	0	1	1	0	0	0	1
ㅁ	11B7	0	0	0	1	0	0	0	1	1	0	1	1	0	1	1	1
ㅂ	11B8	0	0	0	1	0	0	0	1	1	0	1	1	1	0	0	0
ㅄ	11B9	0	0	0	1	0	0	0	1	1	0	1	1	1	0	0	1
ㅅ	11BA	0	0	0	1	0	0	0	1	1	0	1	1	1	0	1	0
ㅆ	11BB	0	0	0	1	0	0	0	1	1	0	1	1	1	0	1	1
ㅇ	11BC	0	0	0	1	0	0	0	1	1	0	1	1	1	1	0	0
ㅈ	11BD	0	0	0	1	0	0	0	1	1	0	1	1	1	1	0	1
ㅊ	11BE	0	0	0	1	0	0	0	1	1	0	1	1	1	1	1	0
ㅋ	11BF	0	0	0	1	0	0	0	1	1	0	1	1	1	1	1	1
ㅌ	11C0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	0
ㅍ	11C1	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	1
ㅎ	11C2	0	0	0	1	0	0	0	1	1	1	0	0	0	0	1	0

: 표 1 위배 부호 : 표 2 위배 부호

2.6 유니코드와 UTF-8 부호

가. UTF 부호

UTF(Universal Character Set(UCS) Transformation Format)는 유니코드와 ISO코드를 8비트 단위의 부호로 부호화하는 것이다. 참고문헌[19]는 UTF 부호화의 종류에 대하여 보여주고 있다. 현재 UTF-1부호 체계는 현재 거의 사용되지 않는다. 표 5와 표 6은 유

표 5. 유니코드의 UTF-8 부호 변환
Table 5. UTF-8 encoding of unicode

Bits	최종부호점	Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6
7	U+007F	0xxxxxxx					
11	U+07FF	110xxxxx	10xxxxxx				
16	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx			
21	U+1FFFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx		
26	U+3FFFFFF	111110xx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	
31	U+7FFFFFFF	1111110x	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx

표 6. 유니코드의 UTF-8 부호 변환 예
Table 6. Example of UTF-8 encoding of unicode

유니코드-UTF8변환 규칙	1	1	1	0	A	A	A	A	1	0	B	B	B	B	C	C	1	0	C	C	D	D	D	D
예 : 유니코드 16진수 1100	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0

니코드를 UTF부호체계로 변환하는 방법과 그 한 예를 보여주는 것이다.

다. 원천부호화 규칙과 UTF-8 부호

참고문헌 [9]의 표5와 표6의 원천부호화 규칙은 UTF-8 부호에 동일하게 적용할 수 있다. 즉 원천부호화 규칙에 부합되는 유니코드는 UTF-8 부호로 변환되어도 원천부호화 규칙에 그대로 부합된다. 따라서 UTF-8 부호체계에 참고문헌[9]를 적용할 수 있다. 원천부호화 규칙에 따라 UTF-8 부호에서 연속된 네 개 이상의 “0”의 비트가 발생되지 않기 위해서는 다음과 같이 조합되어야 한다. 첫 번째로 표6에서 AAAA의 네 비트는 16진수 0과 1이 오지 않도록 해야 한다. 왜냐하면 원천부호화 규칙에 의하면 16진수 E다음에는 16진수 0과 1의 조합이 제한되고 있다. 그런데 UTF-8은 최상위 16진수가 E부터 시작되고 있음을 알 수 있다. 두 번째로 표 6에서 BBBB의 네개의 비트에는 16진수 0과 1이 오지 않도록 해야 된다. 왜냐하면 원천부호화 규칙에 의하면 16진수 마지막 2비트가 10으로 끝나는 16진수 (2, 6, A, E)에는 16진수 0과 1의 조합이 제한되고 있기 때문이다. 또한 BBBB가 16진수 4가 될 때는 CCCC가 16진수 1,2, 3이 오지 않도록 해야 한다. 그리고 BBBB가 16진수 4가 될 때는 CCCC가 16진수 1, 2, 3, 4, 5, 6, 7과 조합되지 않도록 해야한다.

세 번째 비트열 CCCC에 4 또는 C가 될 때는 네 번째 비트열 DDDD에 16진수 1, 2, 3과 조합되지 않도록 해야 한다. 세 번째 비트열 CCCC가 16진수 8이 될 때는 네번째 비트열 DDDD가 16진수 1,2,3,4,5,6,7이 되지 않도록 해야 한다. 또한 세 번째 비트열 CCCC가 16진수 2, 6, A, E가 되면 네 번째 비트열 DDDD가 1로 조합되는 것을 피해야 한다. 이것은 참고문헌 [9]에서 제시한 표 5와 표 6의 원천부호화 규칙과 동일한 규칙임을 알 수 있다.

라. 웹페이지 UTF-8 부호 사용현황

표 7은 인터넷 웹페이지에서 사용되고 있는 각종 부호화 방식들에 대한 적용현황이다. UTF-8 부호화 방식에 의한 웹페이지가 전체의 약68를 점유하고 있다. 그림 4[20][21][22]는 UTF-8부호화 방식과 ISO/IEC 8859 부호화 방식의 웹페이지 적용 추세이다. 이 그림에서 볼 때 UTF-8부호화 방식의 적용은 상승추세인 반면 ISO/IEC 부호화 방식의 적용은 하강추세임을 알 수 있다.

표 7. 원천부호화 방식의 웹사이트 적용 현황
Table 7. Website application of source coding systems

부호화방식	웹사이트수	점유율()
UTF-8	29,744,253	67.52
ISO/IEC 8859	12,191,940	27.68
Shift JIS	954,090	2.17
GB 2312	946,827	2.15
GBK	98,051	0.22
Big5	60,180	0.14
Windows-1256	52,892	0.12
UTF-16/UCS-2	4,444	0.01

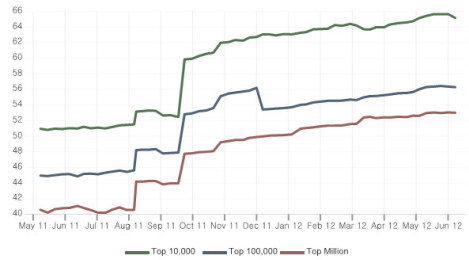

그림 4. UTF-8 부호체계의 웹페이지 적용현황
Fig. 4 Application trend to webpage of UTF-8 coding system

2.7 유니코드 한글낱자의 UTF-8부호

표 8은 이러한 규칙에 따라 유니코드 한글낱자 부호를 UTF-8 부호로 변환한 것을 나타내는 것이다. 이 표 8을 분석해 본 결과 참고문헌[9]의 원천부호화 규칙에 위배되는 부호는 총116개로 나타났다. 또한 표 3에 의하여 초성, 중성 및 종성별 사용 빈도율에 의해 구체적으로 분석해 보면 초성, 중성 및 종성 총 48개의 부호 중에서 원천 부호화 규칙에 위배되는 부호는 116개에 해당된다. 여기서 초성, 중성 및 종성 48개의 모든 낱자에 연속 네 개 이상의 "0"의 비트열이 동일하게 두 개씩 총 102개가 발생한다. 또한 초성에서 ㄱ ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ 및 ㅌ에 추가로 한 개가 더 발생한다. 중성에서"ㅏ"에 추가로 한 개 더 발생하고 종성에서는 ㅋ ㅌ ㅍ ㅎ 및 "ㄹㄱ"에 각각 한 개씩 추가로 발생한다. 표 8은 유니코드와 UTF-8 부호에서 발생하는 원천부호화 규칙에 위배되는 부호수와 위배율을 보여 주고 있다. 아울러 유니코드를 UTF-8 부호화함에 따른 원천부호화 규칙에 위배되는 율에 대

표 8. 유니코드 한글낱자의 UTF-8 부호표
Table 8. UTF-8 code of unicode Hangul Jamo

문자	유니코드	UTF-8 부호																						
ㄱ	1100	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0
ㄲ	1101	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1
ㄴ	1102	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0
ㄷ	1103	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1
ㄸ	1104	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
ㄹ	1105	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	1
ㅁ	1106	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	1	0
ㅂ	1107	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	1	1
ㅃ	1108	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0
ㅅ	1109	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	1
ㅆ	110A	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0
ㅇ	110B	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	1	0	1	1
ㅈ	110C	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	1	1	0	0
ㅊ	110E	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	1	1	1	0
ㅋ	110F	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	1	1	1	1
ㅌ	1110	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0
ㅍ	1111	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	0	1
ㅎ	1112	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0
ㅏ	1161	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	1
ㅐ	1162	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	0	0	0	1	0
ㅑ	1163	1	1	1	0	0	0	0	1	1					1	0	1	1	0	1	0	0	0	1	1
ㅓ	1165	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	0	0	1	0	1
ㅔ	1166	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	0	0	1	1	0
ㅕ	1167	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	0	0	1	1	1
ㅖ	1168	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	0	1	0	0	0
ㅗ	1169	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	0	1	0	0	1
ㅛ	116D	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	0	1	1	0	1
ㅜ	116E	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	0	1	1	1	0
ㅠ	1172	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	1	0	0	1	0

ᅳ	1173	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	1	0	0	1	1
ᅵ	1175	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0	1	1	0	1	1	0	1	0	1
ᆨ	11A8	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	0	1	0	0	0
ᆫ	11AB	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	0	1	1	0	1
ᆬ	11AC	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	0	1	1	0	0
ᆭ	11AD	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	0	1	1	0	1
ᆮ	11AE	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	0	1	1	1	0
ᆯ	11AF	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	0	1	1	1	1
ᆰ	11B0	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	1	0	0	0	0
ᆱ	11B1	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	1	0	0	0	1
ᆷ	11B7	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	1		1	1	1
ᆸ	11B8	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	1	1	0	0	0
ᆹ	11B9	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	1	1	0	0	1
ᆺ	11BA	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	1	1	0	1	0
ᆻ	11BB	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	1	1	0	1	1
ᆼ	11BC	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	1	1	1	0	0
ᆽ	11BD	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	1	1	1	0	1
ᆾ	11BE	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	1	1	1	1	0
ᆿ	11BF	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	1	1	1	1	1	1
ᇀ	11C0	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0
ᇁ	11C1	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	1
ᇂ	11C2	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	1	1	0	0	0	0	0	1	0

: 표 1 위배 부호 : 표 2 위배 부호

하여 보여주고 있다. 결과적으로 유니코드 한글낱자의 경우에 유니코드 자체에서도 원천부호화 규칙에 위배되는 문자 부호가 다수 존재한다. 또한 유니코드를 UTF-8 부호화했을 때, 유니코드 자체에서 보다 초성은 2.7배 중성은 29배, 종성은 11배 원천부호화에 위배되는 부호가 더 발생하는 것으로 나타났다.

III.유니코드 호환용 한글낱자와 원천부호화

3.1 유니코드호환용한글낱자36개 사용빈도율

표 10은 표 5로부터 추출한 36개의 한글 자음과 모음에 대한 사용 빈도율이다.

표 9. Unicode 및 UTF-8 한글낱자 원천부호화 규칙 적합성 비교
Table 9. Compatibility of unicode and UTF-8 coding to source coding rule

구분	한글 낱자	유니코드		UTF-8		(UTF-8/유니코드)위배율(배)
	문자수	원천부호화 위배수	유니코드 위배율(배) (위배수/문자수)	원천부호화 위배수	UTF-8 위배율(배) (위배수/문자수)	
초성	17	16	0.9	43	2.5	2.78
중성	11	1	0.1	29	2.6	26
종성	20	4	0.2	44	2.2	11
계	48	21	0.4	116	2.4	6.0

표 10. 한글 192개 단어의 낱자 사용 빈도율
Table 10. Using rate of Hangul Jamo in mostly using 192 words

낱자	빈도	낱자	빈도
ㅏ	17.83%	ㅋ	1.12%
ㄷ	10.41%	ㄸ	0.79%
ㅇ	9.97%	ㅔ	0.56%
ㄱ	6.89%	ㅌ	0.48%
ㅣ	6.25%	ㅊ	0.46%
ㄴ	5.37%	ㄴㅎ	0.43%
ㅓ	4.41%	ㅍ	0.41%
ㅗ	4.08%	ㅂㅅ	0.34%
ㄹ	4.04%	ㅖ	0.26%
ㅅ	3.51%	ㅛ	0.23%
ㅎ	3.42%	ㅋ	0.22%
ㅈ	3.38%	ㄲ	0.09%
ㅜ	3.32%	ㄹㄱ	0.07%
ㅡ	3.28%	ㅃ	0.07%
ㅁ	3.13%	ㅠ	0.06%
ㅂ	2.21%	ㅑ	0.04%
ㅐ	1.50%	ㄴㅈ	0.03%
ㅆ	1.39%	ㄹㅁ	0.03%

3.2 유니코드 호환용한글낱자 36개와 원천부호화 규칙

표 11은 표 2의 유니코드의 호환용한글낱자(호환용한글자모) 부호중에서 “2.5”에서와 같이 국립국어원에서 조사한 총58,437개의 단어 중에서 사용빈도수가 1000번 이상되는 192개의 단어에 사용된 낱자들에 대한 것이다. 유니코드 호환용한글낱자는 총96개이다. 이중에서 192개의 단어에서 사용된 낱자는 표 10에서와 같이 총 34개이다. 이 표에서 보는 바와 같이 호환용한글 낱자의 부호는 초성, 중성 및 종성으로 구분하고 있지 않다. 자음과 모음으로 구분하여 부호화하고 있다. 표 11에서 총34개의 자음과 모음중에서 참고문헌[9]의 원천부호화 규칙에 위배되는 부호는 ㅁ ㅂ ㅃ의 자음 세 개, ㅐ ㅠ ㅡ의 모음 세 개로서 6개이다. 이 표 11에서 사선표시된 부호는 참고문헌[9] 표6의 원천부호화 규칙에 위배되는 부호로서 6개이다. 그러므로 총 34개의 부호중에서 12개가 원천부호화 규칙에 위배되는 것으로 나타났다. 이 6개의 자음과 모음에 대한 사용 빈도율을 표 11에 의하여 계산하면 참고문헌

표 11. 유니코드 호환용 한글 낱자 34개 부호표
Table 11. Unicode Hangul compatibility Jamo 34ea code table

문자	hexa	유니코드 호환용 한글 낱자 부호표															
ㄱ	3131	0	0	1	1	0	0	0	1	0	0	1	1	0	0	0	1
ㄲ	3132	0	0	1	1	0	0	0	1	0	0	1	1	0	0	1	0
ㄴ	3134	0	0	1	1	0	0	0	1	0	0	1	1	0	1	0	0
ㄴㅈ	3135	0	0	1	1	0	0	0	1	0	0	1	1	0	1	0	1
ㄴㅎ	3136	0	0	1	1	0	0	0	1	0	0	1	1	0	1	1	0
ㄷ	3137	0	0	1	1	0	0	0	1	0	0	1	1	0	1	1	1
ㄸ	3138	0	0	1	1	0	0	0	1	0	0	1	1	1	0	0	0
ㄹ	3139	0	0	1	1	0	0	0	1	0	0	1	1	1	0	0	1
ㄹㄱ	313A	0	0	1	1	0	0	0	1	0	0	1	1	1	0	1	0
ㄹㅁ	313B	0	0	1	1	0	0	0	1	0	0	1	1	1	0	1	1
ㅁ	3141	0	0	1	1	0	0	0	1	0	1	0	0	0	0	0	1
ㅂ	3142	0	0	1	1	0	0	0	1	0	1	0	0	0	0	1	0
ㅃ	3143	0	0	1	1	0	0	0	1	0	1	0	0	0	0	1	1
ㅄ	3144	0	0	1	1	0	0	0	1	0	1	0	0	0	1	0	0
ㅅ	3145	0	0	1	1	0	0	0	1	0	1	0	0	0	1	0	1
ㅆ	3146	0	0	1	1	0	0	0	1	0	1	0	0	0	1	1	0
ㅇ	3147	0	0	1	1	0	0	0	1	0	1	0	0	0	1	1	1
ㅈ	3148	0	0	1	1	0	0	0	1	0	1	0	0	1	0	0	0
ㅊ	314A	0	0	1	1	0	0	0	1	0	1	0	0	1	0	1	0
ㅋ	314B	0	0	1	1	0	0	0	1	0	1	0	0	1	0	1	1
ㅌ	314C	0	0	1	1	0	0	0	1	0	1	0	0	1	1	0	0

ㅍ	314D	0	0	1	1	0	0	0	1	0	1	0	0	1	1	0	1
ㅎ	314E	0	0	1	1	0	0	0	1	0	1	0	0	1	1	1	0
ㅏ	314F	0	0	1	1	0	0	0	1	0	1	0	0	1	1	1	1
ㅐ	3150	0	0	1	1	0	0	0	1	0	1	0	1	0	0	0	0
ㅑ	3151	0	0	1	1	0	0	0	1	0	1	0	1	0	0	0	1
ㅓ	3153	0	0	1	1	0	0	0	1	0	1	0	1	0	0	1	1
ㅔ	3154	0	0	1	1	0	0	0	1	0	1	0	1	0	1	0	0
ㅕ	3155	0	0	1	1	0	0	0	1	0	1	0	1	0	1	0	1
ㅖ	3156	0	0	1	1	0	0	0	1	0	1	0	1	0	1	1	0
ㅗ	3157	0	0	1	1	0	0	0	1	0	1	0	1	0	1	1	1
ㅛ	315B	0	0	1	1	0	0	0	1	0	1	0	1	1	0	1	1
ㅜ	315C	0	0	1	1	0	0	0	1	0	1	0	1	1	1	0	0
ㅠ	3160	0	0	1	1	0	0	0	1	0	1	1	0	0	0	0	0
ㅡ	3161	0	0	1	1	0	0	0	1	0	1	1	0	0	0	0	1
ㅣ	3163	0	0	1	1	0	0	0	1	0	1	1	0	0	0	1	1

: 표 1 위배 부호 : 표 2 위배 부호

[9]표5의 원천부호화 규칙에 대한 위배율은 26.83, 표 6의 원천부화 규칙에 대한 위배율은 14.74, 합계 41.57에 달하는 것으로 나타났다.

표 12는 유니코드 호환용 한글 낱자 부호를 UTF-8로 변환한 UTF-8부호표이다. 이 표 12의 UTF-8 부호도 표 6과 같은 방법으로 유니코드를 UTF-8로 부호화한 것이다. 표 7과 표 13은 자음과 모음에 부여하는 BMP 부호판이 다를 뿐이다. 이 표 12에서 보듯

표 12. 유니코드 호환용 한글 낱자 34개의 UTF-8 부호표
Table 12. UTF-8 code table of unicode Hangul compatibility Jamo 34ea

문자	hexa	유니코드 호환용 한글 낱자 UTF-8 부호																						
ㄱ	3131	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	1	1	0	0	0	1
ㄲ	3132	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	1	1	0	0	1	0
ㄴ	3134	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	1	1	0	1	0	0
ㄴㅈ	3135	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	1	1	0	1	0	1
ㄴㅎ	3136	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	1	1	0	1	1	0
ㄷ	3137	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	1	1	0	1	1	1
ㄸ	3138	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	1	1	1	0	0	0
ㄹ	3139	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	1	1	1	0	0	1
ㄺ	313A	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	1	1	1	0	1	0
ㄻ	313B	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	0	1	0	1	1	1	0	1	1
ㅁ	3141	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	0	0	0	1
ㅂ	3142	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	0	0	1	0
ㅃ	3143	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	0	0	1	1
ㅂㅅ	3144	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	0	1	0	0
ㅅ	3145	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	0	1	0	1
ㅆ	3146	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	0	1	1	0
ㅇ	3147	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	0	1	1	1
ㅈ	3148	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	1	0	0	0
ㅊ	314A	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	1	0	1	0
ㅋ	314B	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	1	0	1	1
ㅌ	314C	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	1	1	0	0

ㅍ	314D	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	1	1	0	1
ㅎ	314E	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	1	1	1	0
ㅏ	314F	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	0	1	1	1	1
ㅐ	3150	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	1	0	0	0	0
ㅑ	3151	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	1	0	0	0	1
ㅓ	3153	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	1	0	0	1	1
ㅔ	3154	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	1	0	1	0	0
ㅕ	3155	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	1	0	1	0	1
ㅖ	3156	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	1	0	1	1	0
ㅗ	3157	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	1	0	1	1	1
ㅛ	315B	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	1	1	0	1	1
ㅜ	315C	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	0	1	1	1	0	0
ㅠ	3160	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0
ㅡ	3161	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	1
ㅣ	3163	1	1	1	0					1	0	0	0	0	1	0	1	1	0	1	0	0	0	1	1

: 표 1 위배 부호 : 표 2 위배 부호

이 호환용 한글 낱자 유니코드를 UTF-8 부호화 함에 따라 참고문헌[9] 표 5의 원천부호화 규칙에 위배되는 부호가 44개로 나타났다. 또한 참고문헌[9] 표 6의 원천부호화 규칙에 위배되는 부호는 6개로 나타났다. 따라서 총 50개의 원천부호화 규칙에 위배되는 부호가 있는 것으로 나타났다. 이것을 표 12에 의하여 원천부호화 규칙에 위배되는 부호들의 사용 빈도율을 계산하면 참고문헌[9] 표 5의 원천부호화 규칙에 대한 위배율은 125.64, 참고문헌[9] 표 6의 원천부화 규칙에 대한 위배율은 13.82, 합계 139.46에 달하는 것으로 나타났다.

표 13은 유니코드 호환용 한글 낱자 부호와 이 부호를 UTF-8 부호화 했을 때 참고문헌[9] 표 5와 표 6의 원천부호화 규칙에 얼마나 위배되는지를 보여 주는 것이다. 유니코드에서는 전체의 36%정도의 부호가 원천 부호화 규칙에 위배된다. UTF-8부호에서는 전체의 150%배정도의 원천부호화 규칙에 위배되는 부호가 있는 것으로 나타났다. 한편 유니코드를 UTF-8 방식으로 부호화함으로서 발생하는 원천부호화 규칙 위배율은 유니코드에 비하여 UTF-8 부호의 위배율

표 13. Unicode 호환용 한글 낱자와 UTF-8 부호의 원천천부호화 규칙 적합성 비교
Table 13. Aptness of unicode Hangul compatibility Jamo and its UTF-8 code for source coding rule

구 분	유니코드	UTF-8
한글낱자수(개)	36	36
원천부호화 위배수(개)	13	53
(위배수/낱자수)(배)	0.36	1.5
(UTF-8/유니코드) 위배율(배)	4.2	
원천부호화규칙 위배낱자 사용 빈도율		
참고문헌[9] 표 5 위배 빈도율(%)	26.83	125.64
참고문헌[9] 표 6 위배 빈도율(%)	14.74	13.82
위배낱자 빈도율 합계(%)	41.57	139.46
(UTF-8/ 유니코드) 위배 낱자 사용 빈도율(%)	335.48	

이 420%정도 증가하는 것으로 나타났다.

Ⅳ. 결 론

본 논문은 유니코드의 한글낱자 및 호환용 한글 낱자와 이 유니코드에 대한 UTF-8 부호가 HDB-3방식의 스크램블링 회선부호화에 적절한지 여부를 [ ]에서 제시한 표 1과 표 2의 원천부호화 규칙에 적합한지 여부를 분석하였다. 연구결과 현재의 유니코드 한글낱자 및 호환용 한글 낱자와 특히 이 유니코드를 UTF-8로 변환한 부호에는 원천부호화 규칙에 위배되는 부호가 상당히 많은 것으로 분석되었다.

특히 유니코드를 UTF-8로 변환한 경우에 원천부호화 규칙에 위배되는 부호가 급격히 증가하는 것으로 나타났다. 유니코드 한글낱자의 경우에는 낱자 48개 중에 21개가 원천부호화 규칙에 위배되었다. UTF-8 한글낱자의 경우에는 낱자 48개를 포함하여 낱자들에 중복하여 나타나는 위배수가 총 116개였다. 또한 UTF-8로 유니코드 한글낱자로 변환한 경우에 원천부호화에 위배되는 부호가 5.75로 증가하였다. 유니코드 호환용 한글 낱자의 경우에는 유니코드의 경우에 한글낱자 34개 중에서 12개 부호가 원천부호화 규칙에 맞지 않는 것으로 나타났다. UTF-8 호환용 한글낱자 부호의 경우에는 한글낱자 부호수 34개를 포함하여 이들 낱자 부호에 중복하여 합계 50개의 원천부호화 규칙 위배 부호가 있는 것으로 분석되었다. 유니코드를 UTF-8로 변환하였을 때 그 위배율이 4.2배 증가하였다.

한글낱자의 사용 빈도율로 분석한 결과 유니코드 한글낱자의 경우에는 원천부호화 규칙에 위배되는 부호에 대한 해당낱자가 포함된 단어의 사용빈도율이 초성은 94, 중성은 42.12 그리고 종성은 2가 되는 것으로 나타났다. 유니코드 한글 낱자의 UTF-9 부호의 경우에는 원천부호화 위배 부호에 대한 사용 빈도율이 초성 258, 중성 242 및 종성 207로 나타났다. 유니코드 호환용 한글 낱자의 경우에는 원천부호화 규칙에 위배되는 부호의 사용 빈도율이 41.57로 나타났다. 또한 유니코드 호환용 한글 낱자의 UTF-8 부호의 경우에는 원천부호화 규칙에 위배되는 부호의 사용 빈도율이 139.46로 나타났다. 도율이 41.57로 나타났다. 또한 유니코드 호환용 한글 낱자의 UTF-8 부호의 경우에는 원천부호화 규칙에 위배되는 부호의 사용 빈도율이 139.46로 나타났다. 본 논문에서는 유니코드와 UTF-8 한글낱자와 한글 호환용 한글낱자의 원천부호가 원천부호화 규칙에 얼마나 부합되는지를 분석하였다. 유니코드 한글낱자 부호와 유니코드 호환용 한글 낱자의 부호는 UTF-8부호로 변환시 가장 좋지 않은 부호11XX와 31XX로 구성되어 있어서 원천부호화 위배율이 더욱 큰 것으로 나타났다. 그러므로 본 논문의 연구결과를 토대로 하여 향후에 본 연구결과로 나타난 문제점을 해결하는 방안을 연구하여야 할 것이다.

감사의 글

이 논문은 2012년도 한세대학교 교내 학술연구비지원에 의해 연구되었음

참고 문헌

[1] Behrouz A. Forouzan, “Data communications" McGraw Hill Korea, pp132-134. Jan, 2008.
[2] http://en.wikipedia.org/wiki/HDB3#B8ZS_.28North_American_T1.29
[3] http://en.wikipedia.org/wiki/HDB3#HDB3_.28European_E-carrier.29
[4] http://searchnetworking.techtarget.com/definition/B8ZS
[5] ITU-T, "Physical/electrical characteristics of hierarchical digital interfaces", ITU-T Recommendation G.703, 1998.
[6] TTA,“Test Method for Telecommunication Terminal Equipment”, TTA Standard TTAS. KO-05.0028/R1, Revised on 23, 2004.
[7] 김용연, “프랙털 영상 부호화에 관한 연구”, 한국전자통신학회논문지, 7권, 3호, pp. 560-565, 2012.
[8] 홍완표, “데이터 전송효율을 고려한 3x4비트 1바이트 문자부호화 규칙에 관한 연구”, 한국전자통신학회논문지, 6권, 4호, pp. 499-504, 2011.
[9] 홍완표, “데이터 전송 효율을 고려한 4비트행x4비트열 2 바이트 문자 부호화 규칙에 관한 연구”, 한국항행학회 논문지, 15권, 5호, 2011.
[10] 한영오, "GRNN을 이용한 동영상 움직임 예측

및 대역분할부호화에 관한 연구", 한국전자통신학회논문지, 5권, 3호, pp. 255-265, 2010.
[11] http://en.wikipedia.org/wiki/Unicode_Transformation_Format#Unicode_Transformation_Format_and_Universal_Character_Set
[12] http://trends.builtwith.com/encoding/UTF-8
[13] http://en.wikipedia.org/wiki/UTF-8
[14] http://en.wikibooks.org/wiki/Unicode/Character_reference/1000-1FFF
[15] http://en.wikibooks.org/wiki/Unicode/Character_reference/3000-3FFF
[16] http://en.wikibooks.org/wiki/Unicode/Versions#Unicode_1.1
[17] http://en.wikibooks.org/wiki/Unicode/Versions#Unicode_1.0
[18] 유진우, 임형규, 박세원, "프랙털 영상 부호화에 관한 연구", 한국전자통신학회논문지, 6권, 2호, pp. 194-197, 2011.
[19] http://en.wikipedia.org/wiki/UTF-1
[20] http://trends.builtwith.com/encoding/UTF-8
[21] http://trends.builtwith.com/encoding/ISO-IEC-8859
[22] http://trends.builtwith.com/encoding/UTF-8
[23] 이창영, "윈도우가 적용된 자기상관에 의한 선형예측부호의 개선선", 한국전자통신학회논문지, 6권, 2호, pp. 186-190, 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 한국정보통신기술사협회장
2002년~현재 한세대학교 정보통신공학과 교수
※ 관심분야 : 위성통신방송, 문자코딩, 통신정책