논문-01-6-1-11

디지털 방송용 한글 허프만 부호 설계 및 PSIP 구조

황 재 정*, 진 경 식*, 한 학 수*, 최 준 영**, 이 진 환**

Huffman Code Design and PSIP Structure of Hangul Data for Digital Broadcasting

Jae-Jeong Hwang*, Kyong-Sig Jin*, Hak-Su Han*, Joon-Young Choi** and Jin-Whan Lee**

요 약

본 논문은 한글 텍스트 데이터에 대한 부호화 효율을 극대화시키는 관점에서 예외 부호화를 통해 최적의 허프만 부호를 얻는다. 한글 코드는 표준 완성형과 유니 코드를 대상으로 하였으며 같은 허프만 부호를 부여하였다. 현재 우리나라의 디지털 TV는 한글 문자를 압축하지 않고 전송하는 형태이며, 본격적인 데이터 방송이 시작되면 한글 데이터가 차지하는 전송량으로 인한 심각한 문제가 야기된다. 본 논문에서는 데이터 방송에서 문제가 되는 전송량을 줄이기 위해 한글 전용 최적의 허프만 부호를 생성한다. 미국의ATSC 표준을 바탕으로 한 디지털 TV 국내 표준에 적용하기 위해 프로그램 및 시스템 프로토콜(PSIP) 구조를 제안한다. 결과로서, 발생확률 0.0043 이하의 확률을 갖는 문자를 예외 부호화하여 최대 46%의 압축율을 얻는 기법을 제안하였다.

Abstract

In this paper, we derive an optimal Huffman code set with escape coding that maximizes coding efficiency for the Hangul text data. The Hangul code can be represented in the standard Wansung or Unicode format, and we can generate a set of Huffman codes for both. The current Korean DTV standard has not defined a Hangul compression algorithm which may be confronted with a serious data rate for the digital data broadcasting system. Generation of the optimal Huffman code set is to solve the data transmission problem. A relevant PSIP structure for the DTV standard is also proposed. As a result, characters which have the probability of less than 0.0043 are escape coded, showing the optimum compression efficiency of 46%.

I. 서 론

1940년대부터 TV 방송이 시작된 이래 아날로그 시대를 거쳐 디지털 TV에 대한 연구가 1990년대 들어 활발히 전개되었으며 ATSC(Advanced Television Systems Committee) 방식으로 일컬어지는 미국식 시스템과 DVB(Digital Video Broadcasting)라 명명된 유럽식 DTV(Digital Television)가 개발되었다. 우리나라에서 채택한 DTV 표준은 ATSC 방식을 근간으로 하였으며[1], 최대 19.4 Mbps 범위내에서 영상, 음향, 데이터 등 정보를 필요에 따라 전송할 수 있다. 선명한 화질을 가시청 지역내에서 보장하고 HDTV를 전송할 수 있다는 디지털 TV의 특성이 있으나 시청자의 욕구는 데이터 방송과 같은 양방향 서비스를 선호하고 새로운 서비스가 제공되어야만 DTV는 성공적으로 정착될 수 있을 것이다[2]. 다중의 서비스, 고속 데이터 전송, 인터넷 서비스 등 프로그램과 관련/비관련한 다양한 기능을 제공하는 데이터 방송의 정보량은 최대 19.4Mbps까지 확대 가능하므로 압축의 필요성이 대두되는 것이다[3][4][5]. 아날로그 TV에서 자막 방송은 데이터량이 크지 않아 9600bps

* 군산대학교 전자정보공학부
School of Electronic & Information Eng., Kunsan National Univ.
** 한국전자통신연구원
Radio & Broadcasting Tech. Lab., ETRI
※ 이 논문은 한국전자통신연구원의 지원으로 이루어졌음

정도의 속도로 충분하였으나[6] 데이터 방송에서는 영상 데이터 만큼의 정보도 전송될 필요가 있는 것이다.

한편 미국의 ATSC에서는 영문 텍스트 데이터 전송을 위해 PSIP(Program and System Information Protocol : A/65) 규격을 제정하였다[7]. 이 표준에서는 영문자에 대한 압축으로 허프만 코드를 사용하고 있으며 또한 타이틀과 내용에 따라 별도의 테이블을 사용하고 있다. 허프만 코드는 2차로 정의되며 발생 빈도가 적은 알파벳은 예외 부호화된다[8][9]. 그러나 국내 디지털 방송 규격[10]에서는 한글코드 압축 테이블이 마련되어 있지 않아 현재까지 한글문자를 압축하지 않고 전송하는 형태를 취하고 있어 2001년부터 디지털 방송이 시작되고 데이터 방송이 본격화되면 한글 데이터가 차지하는 전송량으로 인해 심각한 문제가 야기될 것으로 예상된다.

따라서 본 논문에서는 영문에 대한 허프만 압축 기술을 응용하여 한글에 맞게 적용하도록 한다. 먼저 한글 코드의 특징을 고찰하고 표준 조합형과 표준 완성형[11][12][13] 및 유니코드[14]를 비교 분석하며, 예외 부호화를 할 때, 각각의 압축율과 테이블 용량면에서 서로 비교 분석을 통해 부호화 효율이 가장 높은 곳에서 허프만 코드를 생성하도록 한다. PSIP의 텍스트 압축 관련 표준 분석을 통해 표준완성형코드를 기준으로 허프만 부호화를 이용한 한글 압축 코드를 제안한다.

제2장에서 현재 사용하는 한글 코드의 종류와 구조에 대해 살펴보고, 제3장에서는 우리나라에서 채택하고 있는 DTV 규격에서 텍스트 데이터 처리 부분을 검토하고 제안한 한글 허프만 코드 발생 기법을 설명한다. 제4장에서는 PSIP에 한글 허프만 코드를 적용하는 문제를 다룬다.

Ⅱ. 한글 코드 종류 및 구조

현재의 한글 코드는 두 가지(표준 조합형과 표준 완성형)[11][12][13] 모두를 사용하고 있으며 수시로 변환이 필요한 경우가 많다. 내부 코드 처리를 위해서는 비교적 간단한 조합형을 이용하고 외부 출력을 위해서는 자체의 변환 엔진을 이용하여 완성형을 사용한다. 또한 조합형과 완성형에 관한 영역을 같이 배정하고 있는 유니코드도 사용한다.

2.1 표준 조합형 코드(KS X 1001 : KS C 5601-1992)

자소에 의미를 부여하여 초성, 중성, 종성으로써 가능한 한글(11,172자)을 모두 표현할 수 있으며 자소 당 5비트를 할당한다. 코드 체계는 그림 1과 같으며 2 바이트의 하위부터 5비트씩 종성, 중성, 초성에 할당하고 MSB를 1로 세트한다.

Second Byte								First Byte							
MSB	7	6	5	4	3	2	1	8	7	6	5	4	3	2	1
1	초성					중성					종성				

그림 1. 표준 조합형 코드 체계
Fig. 1. Standard Johab code system

2.2 표준 완성형 코드(KS X 1001 : KS C 5601-1987)

음절에 의미를 부여하고 일반적으로 자주 쓰이는 한글(2350자)만을 표현하며 음절당 2바이트를 할당한다. 코드 체계는 그림 2와 같으며 각 바이트의 상위 비트를 1로 세트한다. 영문자를 위한 아스키 코드는 7비트로 이루어지므로 한글과 아스키는 구분된다. 반면 조합형 코드는 하위 바이트의 상위 비트가 1 또는 0으로 변하므로 통신 제어 코드와 충돌을 일으킬 수 있는 문제가 있다.

Second Byte								First Byte							
MSB	7	6	5	4	3	2	1	MSB	7	6	5	4	3	2	1
1	Data							1	Data						

그림 2. 표준 완성형 코드 체계
Fig. 2. Standard Wansung code system

2.3 유니코드(KS X 1005-1 : KS C 5700, ISO/IEC 10646-1)

유니코드는 다국어 언어의 효율적인 처리를 위해 1989년 컨소시엄 형태로 활동을 시작하여 1991년 말에 유니코드 1.0을 발표(표준 완성형 코드만 포함)하였고, 1995년 유니코드 2.0에서는 조합형(1100 - 11F9)과 완성형(AC00 - D7A3)의 두 가지 영역으로 나누어 코드를 부여하였다. 완성형과 같이 음절에 의미를 부여하나 조합형에서 가능한 모든 문자에 대한 코드를 배정하고 있으며 한 음절당 2바이트로 표현된다. 코드 체계는 그림 3과 같으며 MSB를 1로 세트하고 나머지 15비트에 한글 코드를 부여한다.

Second Byte								First Byte							
MSB	7	6	5	4	3	2	1	8	7	6	5	4	3	2	1
1	Data														

그림 3. 완성형 유니코드 체계
Fig. 3. Wansung Unicode system

국내 지상파 DTV 잠정규격(안)에서는 완성형과 유니코드를 사용하도록 규정하고 있다. 완성형에서는 영문 95자, 한글 2,350자, 특수 문자 986자, 한자 4,888자를 지원하며, 유니코드에서는 영문 95자와 보충 문자 96자, 한글 11,172자, 특수 문자 986자, 한자 7,744자를 지원하도록 규정하고 있다. 아래 절에서는 국내 TV방송에서 자주 쓰이는 문자(주로 한글과 일부 영문)를 중심으로 허프만 코드를 생성하고 DTV 규격에 적용하는 문제를 다룬다.

Ⅲ. 한글 데이터를 위한 DTV 규격

국내 지상파 DTV 잠정규격(안)은 미국의 ATSC의 표준을 따르고 있다. 따라서 영문 PSIP의 다중 문자열 구조(Multiple String Structure)와 텍스트 압축을 위한 표준 허프만 테이블(Standard Huffman Tables for Text Compression)을 분석한다.

2.1 다중 문자열 구조

다중 문자열 구조는 텍스트 문자열을 나타내는데 사용된다. 텍스트 문자열은 프로그램 제목(event title), 전체 채널명(long channel names), ETT(Extended Text Table) 메시지, RRT(Rating Region Table) 텍스트 항목들을 나타내는데 사용되어진다. 표 1은 다중 문자열 구조의 비트 스트림 구문을 보여준다[7].

number_strings는 8bit unsigned integer로 number_string 정보 이후에 나타날 문자열의 수를 나타내며

표 1. 다중 문자열 구조의 비트 스트림 구문
Table. 1. Bit stream syntax in multiple string structure

Syntax	Bits	Format
multiple_string_structure( ) {		
number_strings	8	uimsbf
for (i=0;i<number_strings;i++) {		
ISO_639_language_code	8 * 3	uimsbf
number_segments	8	uimsbf
for(j=0;j<number_segments;j++) {		
compression_type	8	uimsbf
mode	8	uimsbf
number_bytes	8	uimsbf
for (k=0;k<number_bytes;k++)		
compressed_string_byte[k]	8	bslbf
}		
}		
}		

ISO_639_language_code는 ISO 639.2/B에 기초한 3-byte로서, i번째 문자열을 위해 사용되는 언어 코드를 나타낸다. number_segments는 8-bit unsigned field로 이 필드이후에 나타날 세그먼트의 개수를 나타내고 compression_type은 8bit 필드로 j번째 세그먼트에 대한 압축 형식을 나타낸다. 이 필드에 허용되는 값들은 표 2와 같다.

표 2. 영문 PSIP의 압축 형식
Table 2. Compression type for English PSIP

압축형식	압축 방법
0x00	압축 안함
0x01	ATSC A/65의 부록 C에 있는 표 C.4와 표 C.5에 정의된 허프만 코드 방법
0x02	ATSC A/65의 부록 C에 있는 표 C.6과 표 C.7에 정의된 허프만 코드 방법
0x03 to 0xAF	reserved
0xB0 to 0xFF	user private

mode는 다음에 오는 세그먼트의 문자를 해독하기 위해 사용되는 텍스트 모드를 나타내는 8-bit 값이다. 표 3은 모드 값을 나타내고 있다.

표 3. 영문 PSIP의 모드
Table 3. Mode for English PSIP

값	의 미	언어 및 스크립터
0x00~0x3E	Select 8-bit ISO/IEC 10646-1	각 언어 및 Reserved
0x3F	Select 16-bit ISO/IEC 10646-1	all
0x40~0xDF	Reserved	
oxE0~0xFE	User private	
oxFF	Not applicable	

number_byte는 8-bit unsigned integer field로 number_byte 이후의 compressed_string_byte수를 나타낸다. compressed_string_byte[k]는 j번째 세그먼트의 k번째 바이트를 의미한다. 다중 문자 구조의 비트 스트림을 간단히 표현하면 그림 4와 같이 64비트의 헤더를 전치하고 가변의 압축된 데이터를 첨부하여전송된다.

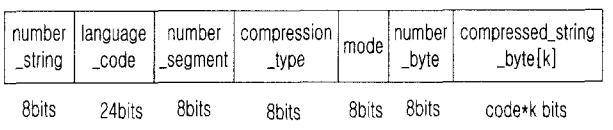

그림 4. 다중 문자열의 전송 구조
Fig. 4. Transmission structure of multiple string

2.2 텍스트 압축을 위한 표준 허프만 테이블

PSIP에서는 텍스트 문자열을 전송하기 위해 텍스트 문자열을 두 가지 형식인 타이틀과 프로그램 내용으로 구분하여 서로 다른 테이블을 할당하였다. 이 허프만 테이블들은 2차(order-1)의 상태적 확률에 기반으로 정의된다. 이들은 각각의 표준 허프만 인코딩 테이블과 디코딩 테이블을 가진다.

2.2.1 허프만 인코딩 테이블

PSIP에서 압축 방법은 ISO/IEC 8859-1 (Latin-1)의 문자를 지원하는데, ASCII 코드(0~127)를 압축한다. 특수한 목적을 사용하도록 종료 문자(Terminate character: ASCII 0)와 예외 문자(Escape character: ASCII 27)에 대해 정의되어 있다. 이들 문자들에 대해 사용의 예외부호화의 과정을 알아보면 그림 5와 같다.

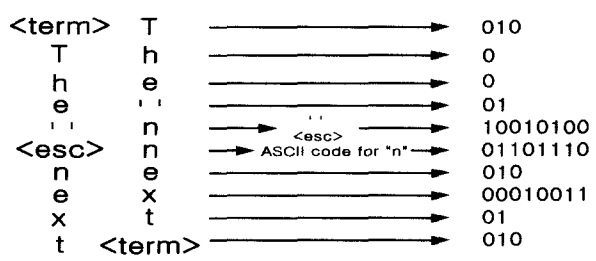

그림 5. 영문 허프만 부호화의 예
Fig. 5. English decoding table format

그림 5에서 종료 문자(term)는 문자열의 처음과 마지막에 위치해 있다. 그 이유는 이 문자의 용도가 문자의 수를 맞추기 위해 채움 문자로 사용되기 때문이다. 이스케이프 문자(esc)는 각각의 코드에 모두 할당되어 있다 이 문자 다음에 오는 문자가 압축되지 않음을 의미한다.

2.2.2 허프만 디코딩 테이블

표 4, 표 5에 나타난 바와 같이 PSIP의 영문 허프만 디코딩 테이블은 2개의 섹션(section)으로 구성된다.

PSIP의 영문은 2차 허프만 부호화로 만들어졌다. 즉, 각 문자들에 대해 트리 루트를 찾고 그 아래에 트리(tree)를 가지게 되는 것이다. 표 4에서 보면 트리를 생성하여 잎(leaf)에 있는 값을 찾는 것이다. 각 문자에 적용되는

표 4. 영문 디코딩 테이블 포맷
Table 4. English decoding table format

Syntax	Bits	Format

decode_table( ) {
for (i= =0; i<128; i++) {
byte_offset_of_char_i_tree_root
}
for (i= =0; i<128; i++) {
character_i_order_1_tree( )
}
}
	16
	
	8*M
		uimsbf

표 5. 영문 디코딩 트리 포맷
Table 5. English decoding tree format

Syntax	Bits	Format

character_i_order_1_tree( ) {
for (j= =0; j<N; j++) {
 	
left_child_word_offset_or_char_leaf
 	
right_child_word_offset_or_char_leaf
}
}
	
	8
	8
		
		uimsbf
		uimsbf

디코딩 트리 포맷을 보면 표 5와 같다.

PSIP의 디코딩 테이블을 보면 양의 정수형 숫자로 쓰여있다. 이유는 디코딩 테이블의 용량을 줄이기 위해서다. 이 테이블은 트리 구조와 문자의 정보가 포함되어 있다.

그림 6은 문자 "x"에 대한 디코딩 테이블을 보이고 있다. 맨 위부터 두 개씩 묶어 각 노드에 대하여 "0"과 "1"을 할당한다. 맨 위는 root가 되고 나머지는 차례로 노드

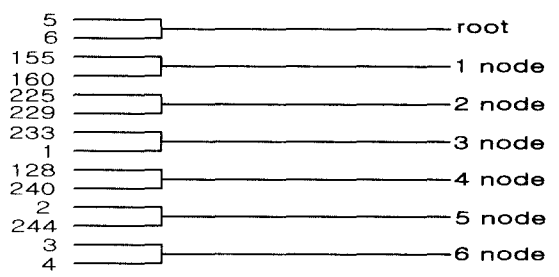

그림 6. 문자 'x'에 대한 디코딩 테이블 예
Fig. 6. Decoding table example for character 'X'

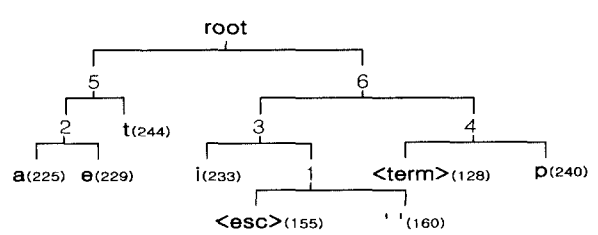

그림 7. 문자 'x'에 대한 트리 구조
Fig. 7. Tree structure for character 'X'

가 정해진다. 여기서 127이하의 값은 노드 번호가 되고 128이상의 수는 문자(ASCII 코드 값)에 128을 더한 수이다. 그림 6을 트리 구조로 나타내면 그림 7과 같다.

Ⅳ. 한글 PSIP의 적용

ATSC의 규격을 근간으로 하는 국내 지상파 DTV 잠정규격(안)에 한글 압축 텍스트 문자열 적용 방법을 제안하며 한글 허프만 부호화 대한 예를 보인다.

3.1 PSIP 포맷 제안

현재까지 지상파 DTV 잠정규격(안)에서 한글에 대한 압축 형식은 아직 할당되어 있지 않다. 우리는 reserved 영역이나 user private 영역에 한글 압축 형식을 적용시킬 수 있다. 국내 지상파 DTV 규격의 압축 형식을 표 6과 같이 제안한다.

표 6. 국내 지상파 DTV규격의 압축 형식 제안
Table 6. Proposed compression type for Korean terrestrial DTV system

압축형식	압축 방법
0x00	압축 안함
0x01	ATSC A/65의 부록 C에 있는 표 C.4와 표 C.5에 정의된 허프만 코드 방법
0x02	ATSC A/65의 부록 C에 있는 표 C.6과 표 C.7에 정의된 허프만 코드 방법
0x03 to 0x9F	reserved
0xA0	유니코드용(KS X 1005-1) 허프만 코드
0xA1 to 0xAF	reserved
0xB0	표준완성형코드용(KS X 1001) 허프만 코드
0xB1	표준조합형코드용(KS X 1001) 허프만 코드
0xB2 to 0xFF	user private

표 7. 국내 지상파 DTV 잠정규격(안)의 모드
Table 7. Modes for Korean terrestrial DTV system

값	의미	언어 or 스크립터
0x00~0x3E	Select 8-bit ISO/IEC 10646-1	각 언어 및 Reserved
0x3F	Select 16-bit ISO/IEC 10646-1	KS X 1005-1 유니코드 한글 (KS C 5700)
0x40~0x47	Reserved	
0x48	표준완성형	KS X 1001 완성형 (KS C 5601)
0x49	표준조합형	KS X 1001 조합형 (KS C 5601)
0x4A~0xE2 0xE3~0xFE	Reserved User private	
0xFF	Not applicable	

언어 모드는 국내 지상파 DTV 잠정규격(안)에서 정의한 표 7과 같이 유니코드, 표준 완성형, 표준 조합형에 대해 0x3F, 0x48, 0x49를 각각 사용한다.

3.2 한글 텍스트의 허프만 테이블

영문 텍스트 문자열에 대한 허프만 테이블은 미국의 ATSC 방식에서 제안된 바가 있다[7]. 영문 텍스트는 알파벳 26자가 대문자와 소문자로 표현되므로 2차 허프만 부호화가 가능하다. 그러나 한글의 경우 영문과 달리 방식에 따라 다르지만 2,350자 이상의 많은 문자수를 가지므로 2차 부호화에서 제곱의 심볼 수가 발생하여 복잡도가 무한히 증가하고 디코딩 테이블 역시 방대해진다. 따라서 아래에 1차(order-0) 한글 허프만 부호화 기법을 제안하였다.

3.2.1 한글 허프만 인코딩 테이블

한글에서 특수용 문자로 이스케이프 문자만 사용키로 한다. 왜냐하면 제안된 한글 허프만 코드는 1차(order-0)로 하기 때문이다.

그림 8은 '똠방각하'라는 문자열에 대한 부호화를 보인

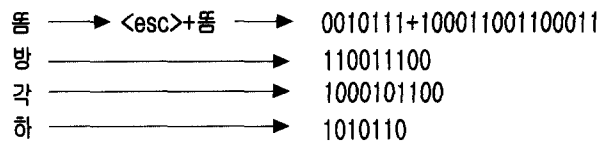

그림 8. 한글 허프만 부호화의 예
Fig. 8. Hangul Huffman coding example

다. '똠'은 표준완성형코드를 벗어난 문자이기 때문에 한글 허프만 부호화에 적용되지 않는다. 이 문자는 예외 부호화를 하는데 있어 한글 허프만 부호화에 적용된 ESC와 '똠'의 확장완성형코드 값을 쓰게 된다.

3.2.2 한글 허프만 디코딩 테이블

영문 PSIP의 허프만 디코딩 포맷에 한글을 적용시키는 문제가 있다. 한글은 1차원 허프만 코딩을 할 경우에 표 5에서 각 문자의 트리 루트가 필요가 없다. 따라서 표 8과 같이 제안한다.

표 8. 국내 DTV 규격의 디코딩 테이블 포맷 제안
Table 8. Proposed decoding table format for Korean terrestrial DTV system

Syntax	Bits	Format
decode_table() {		
for (j= =0; j<N; j++) {		
left_child_word_offset_or_char_leaf	16	uimsbf
right_child_word_offset_or_char_leaf	16	uimsbf
}		
}		

그림 6의 영문 디코딩 테이블에 대응하는 한글 디코딩 테이블의 예를 그림 9에 제시한다. 서로 두 개씩 묶어 위는 상위 바이트, 아래는 하위 바이트를 의미한다.

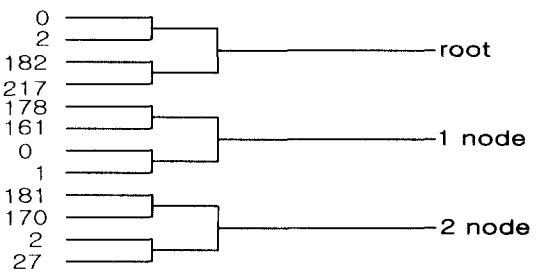

그림 9. 한글 디코딩 테이블 예
Fig. 9.

한글 허프만 디코딩 테이블이 영문의 것과 차이점이 있다면, 한글에서는 노드나 코드가 모두 16 비트로 구성되어 있다는 것이다. 노드와 코드를 구별하기 위해서 코드 값에 부호화 한 문자 수 보다 큰 수를 더하는데 있어 상위 바이트에 1024를 더한다. 그림 9를 트리 구조로 나타내면 그림 10과 같다.

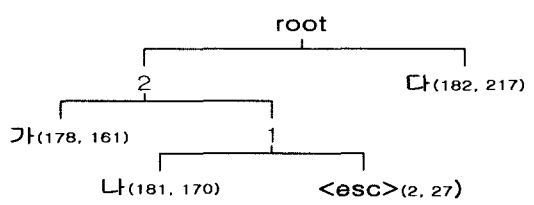

그림 10. 한글 디코딩을 위한 트리 구조 예
Fig. 10. Tree structure example for Hangul decoding

V. 실험결과 및 고찰

디지털 방송에서 발생할 한글데이터는 현재 방송중인 아날로그 방송에서 사용된 데이터와 유사한 것으로 가정하여 표 9에서 보는 것과 같이 KBS, MBC, SBS 등 방송 3사에서 2000. 5월부터 10월까지 6개월간 방송된 뉴스, 드라마, 및 영화 데이터를 수집하였다. 데이터에는 표준완성형 한글 675만자(2바이트/자), ASCII 문자 328만자(1바이트/자), 기타 한자 등 특수문자 2만 2천자(2바이트/자)가 포함되어 있으며 이 중 한글과 영문 데이터만을 실험의 대상으로 하며 기타 문자는 예외 부호로 처리하였다.

표 9. 실험에 사용된 분야별 데이터의 총 문자 수
Table 9. Number of characters used in experiment

구분	뉴스	드라마	영화	합계
표준완성형문자(자)	3,206,678	3,024,132	519,380	6,750,190
ASCII 문자(자)	1,380,927	1,618,493	283,159	3,282,579
기타문자(자)	20,003	2,469	175	22,647
총 용량(byte)	7,834,289	7,671,695	1,322,269	16,828,253

5.1 프로그램별 한글 발생 확률분포

각각의 프로그램별(드라마, 뉴스, 영화)한글의 발생 확률을 서로 비교한다. 각각의 분야별로 발생 확률분포를 보면 거의 비슷한 결과를 볼 수가 있다.

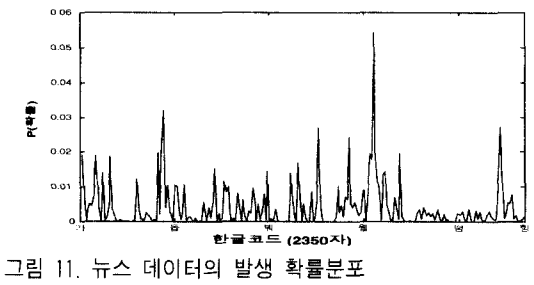

그림 11. 뉴스 데이터의 발생 확률분포
Fig. 11. Probability density of 'news' data

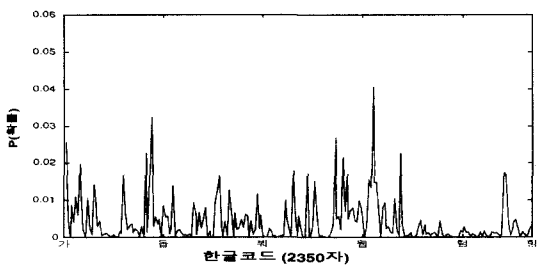

그림 12. 드라마 데이터의 발생 확률분포
Fig. 12. Probability density of 'drama' data

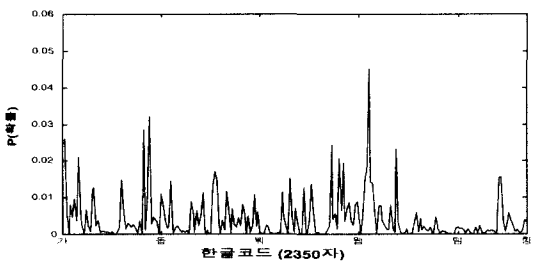

그림 13. 영화 데이터의 발생 확률분포
Fig. 13. Probability density of 'movie' data

5.2 표준 완성형과 유니코드 부호화 비교

표준 완성형과 유니코드에서 한글 발생 확률분포를 비교하면 그림 14 및 그림 15와 같다.

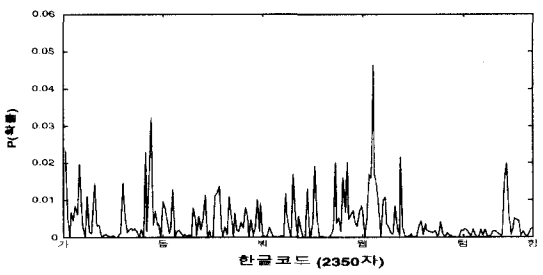

그림 14. 표준완성형코드 발생 확률분포
Fig. 14. Probability density in the standard Wansung code system

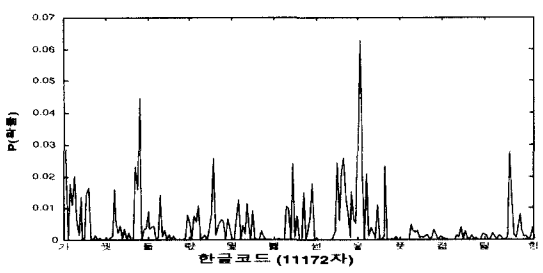

그림 15. 유니코드 발생 확률분포
Fig. 15. Probability density in the Unicode code system

위 실험은 3개의 분야별 집단을 하나로 묶어서 표준완성형은 10자씩, 유니코드는 50자씩 묶어서 비교한 것으로 서로 발생 확률이 매우 유사함을 알 수 있다. 그 이유는 표준 완성형 코드 이외의 한글 문자는 매우 낮은 발생 확률을 가지기 때문이다.

표준 완성형과 유니코드의 허프만 부호화의 비교는 표 10과 같다. 여기서 허프만 평균 부호길이와 평균 부호길이를 분류하는 이유는 한글 문서에는 한글과 아스키 문자 이외의 문자 때문이다.

표 10. 표준 완성형과 유니코드의 허프만 코딩 비교
Table 10. Huffman coding comparison of the standard Wansung and Unicode result

구분	표준 완성형 코드	유니코드
엔트로피	7.195822[b/s]	7.203528[b/s]
허프만 평균 부호길이	7.465641[b/s]	7.497881[b/s]
평균 부호길이	7.520426[b/s]	7.532901[b/s]

따라서 그림 14와 그림 15, 그리고 표 10에서 보는 것과 같이 이 두 코드는 서로 문자코드만 다를 뿐 허프만 부호화에서는 거의 같다고 보면 된다. 따라서 유니코드도 표준 완성형 코드를 위한 허프만 코드를 같이 사용하도록 한다.

5.3 표준 완성형과 표준 조합형 코드 허프만 부호화

표준 완성형은 2바이트를 하나의 샘플로 하여 허프만 부호화를 하였으나, 표준 조합형은 자소 즉, 5bit를 하나의 샘플로 하여 부호화를 한 결과는 표 11과 같다.

위 실험을 통해 표준 조합형 코드인 경우, 자소 당 비트 수와 평균부호길이를 비교하면 약간 줄었음을 볼 수 있

표 11. 표준 완성형 코드와 표준 조합형 코드의 비교
Table 11. Comparison of the standard Wansung and Johab code result

분야	코드 / 비교내용	조합형 코드	완성형 코드
뉴스	허프만 평균 부호길이(b/s)	4.9285	7.1129
	압축율(%)	10.4428	46.6626
드라마	허프만 평균 부호길이(b/s)	4.8635	6.9200
	압축율(%)	12.4392	46.8085
영화	허프만 평균 부호길이(b/s)	4.8650	6.9326
	압축율(%)	13.1723	46.2943

다. 그러나 표준 완성형인 경우 음절당 비트수와 허프만평균 부호길이를 비교할 때 약 8비트가 감소한 것을 볼 수 있다. 또한 압축율에서도 월등하다는 것을 알 수 있다.

5.4 예외 부호화 문자 선정

사상의 수가 증가하면 대응하는 허프만 코드의 수와 코드 길이가 증가하게 되며 구현시 많은 메모리가 필요하게 된다. 따라서 일반적으로 발생 확률이 적은 사상에 대해 코드를 부여하지 않고 예외적으로 부호화한다. 이는 전체 평균 부호화 길이를 증가시키는데 본 실험에서는 최대의 부호화 효율을 가지는 지점에서의 확률적으로 낮은 심볼을 예외 부호화를 통해 실험을 행하였다.

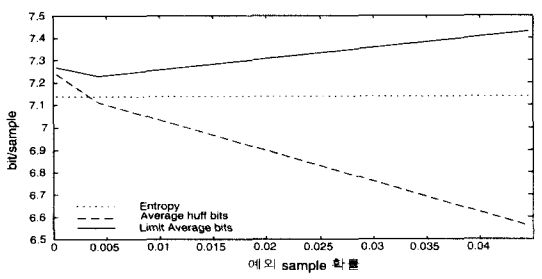

그림 16. 예외 부호화에 따른 평균 부호길이
Fig. 16. Average code length using escape coding

그림 16에서 가장 높은 부호화 효율을 가지는 지점은 엔트로피와 부호화 평균 길이가 차가 가장 적은 지점이 된다. 즉, 그 지점이 이상의 확률을 가지는 문자만 허프만 부호화 적용된다. 전체(발생 된 모든 문자) 허프만 부호화 한 것과 예외(부호화 효율에 의해 선택된 문자) 허프만 부호화 한 것을 비교하면 표 12와 같다.

표 12. 전체와 예외 허프만 부호화 비교
Table 12. Total and escape Huffman coding result

구분 \ data		뉴스	드라마	영화
엔트로피(bits/sample)		7.1129	6.9200	6.9326
전체 부호화	압축율(%)	46.6626	46.8085	46.2943
	평균부호길이(b/s)	7.2551	7.0285	7.0773
	디코딩 테이블 용량(kb)	6.8	7.4	6.6
예외 부호화	압축율(%)	48.9601	49.0704	48.8145
	평균부호길이(b/s)	7.2157	6.9828	7.0084
	디코딩 테이블 용량(kb)	4.4	5.0	4.2

표 12를 보면 예외 허프만 부호화를 한 경우 평균 부호길이는 최대 0.07b/s만큼 감소하며 압축율은 2.5%만큼 증가하여 확률적으로 발생 빈도가 적은 문자에 대해 코드를 부여하지 않고 예외적으로 처리하여 효율을 높인다. 한편 뉴스, 드라마, 영화의 데이터를 모두 통합하여 예외 허프만 부호화한 결과는 표 13과 같다.

표 13. 전체 데이터의 부호화 결과
Table 13. Coding result for total data

구분 \ data	통합 데이터
엔트로피(bits/sample)	7.1378
압축율(%)	45.9938
평균 부호길이(b/s)	7.2306
디코딩 테이블 용량(kb)	3.996

표 13의 결과를 얻기 위한 실험에서 전체 샘플에 대해 0.0043 이하의 확률을 가지는 문자들은 허프만 부호화에서 예외 부호화를 통해 부호화 효율이 가장 높은 허프만 부호를 얻었다. 즉, 한글 전용 허프만 부호화에 적용된 문자 종류는 1000개의 문자이다. 여기서 허프만 부호화에서 적용되지 않는 문자에 대해서는 허프만 부호화 된 ESC코드(ASCII:27)를 붙여 전송함으로써 압축된 데이터와 구분한다. 전체 데이터를 제안한 코드에 의해 부호화한 결과 평균 부호길이는 엔트로피보다 0.1비트 차이로 접근하여 부호화되고 압축율은 약 46%로 나타났다. 압축율은 한글과 영문에 대해 허프만 코드로 부호화하고 기타 문자에 대해 예외부호화한 결과를 총 용량 16.83Mbytes와 비교한 결과이다.

Ⅵ. 결 론

이제까지 살펴본 바와 같이, 표준 완성형과 유니코드의 코드 값만 다를 뿐 압축 부호를 같이 쓸 수 있으며, 발생된 모든 문자에 대해 허프만 부호화를 적용하지 않고 부호화의 효율을 이용하여 발생 확률이 낮은 문자에 대해서는 예외 부호화를 적용시켜 부호화 된 것과 서로 구별한다. 영문의 경우 심볼 수가 적기 때문에 2차 허프만 부호화가 가능하지만, 한글의 경우 심볼 수가 최대 11,172자이며 자주 발생하는 약 1,000개의 문자만 고려해도 1백만개의 심볼이 발생하므로 이 연구에서는 1차 허프만 부호

를 생성하여 디지털 방송용 한글 전용 인코딩 테이블과 디코딩 테이블을 얻어냈다.

조합형, 완성형 및 유니코드를 혼용하는 한글 구조상 서로 다른 허프만 코드를 설계해야 하나, 조합형은 DTV 규격에서도 제외하는 추세이고 실험에 의해 얻어진 바와 같이 완성형과 유니코드는 비슷한 특성을 갖기 때문에 단일의 코드를 제안하였다.

제안한 코드는 원 데이터의 엔트로피에 비해 1.3%만을 초과하는 결과를 보여서 매우 근접한 특성을 갖는다. 원 데이터 16bits/sample에 비한 압축율은 약 55%이나 실제 적용에서 예외 부호화의 영향으로 최종 압축율은 약 46%로 나타났다. 수신기 제작에서 중요한 것은 코드의 수와 메모리 크기이다. 코드의 수가 많으면 메모리가 증가해야 하고 복잡해진다. 여기서는 이 점을 고려하여 메모리 크기 4kb 이하에서 설계하였다. 본 논문에 제안된 방식을 DTV에 적용하여 데이터 방송에서 발생하는 방대한 한글 정보량을 압축하여 전송할 수 있음을 보여준다.

참 고 문 헌

[1] ATSC standard A/53, *ATSC digital television standard*, Sep. 1995.

[2] ATSC Interactive services protocols for terrestrial broadcast and cable, Feb. 1999.

[3] 황재정, 정동훈, "DTV를 위한 데이터 방송 시스템", *대한전자공학회 하계학술대회*, pp. 507-510, Jun. 1999.

[4] ATSC data broadcast specification, Mar. 1999.

[5] http://toocan.philabs.research.philips.com

[6] EIA 708, Specification for advanced television closed captioning (ATVCC), *Electronic Industry Association*, Jul. 1997.

[7] ATSC standard A/65, Program and system information protocol for terrestrial broadcast and cable, Dec. 1997.

[8] K. Sayood, Introduction to data compression, *Morgan Kaufman Publishers*, 2000.

[9] M. Nelson, J. Gailly, The data compression book, M&T books, 1996.

[10] 지상파 디지털 TV 실험방송전단반 서브그룹 2, 지상파 디지털 텔레비젼 방송 규격(안) v10, 2000. 7.

[11] 한국어정보처리연구소, C로 구현한 한글 코드 시스템 프로그래밍, 도서출판 골드, 1999.

[12] 한국전산원, 한글코드 표준적용의 문제사례 분석 기술보고서, 1997. 9.

[13] 국립국어연구원, 한글코드에 관한 연구, 1995.

[14] The Unicode Consortium, The Unicode Standard Version 3.0, Addison Wesley Longman, Inc. 2000.

저 자 소 개

황 재 정

1992년 ~ 현재 : 군산대학교 전자정보공학부 교수
1992년 : 전북대학교 전자공학과 공학박사
1993년 ~ 1994년 : 텍사스주립대학교 객원교수
1990년 ~ 1991년 : 한독기술협력사업 파독교수
2001년 ~ 현재 : 군산대학교 공학연구소장
1997년 ~ 1998년 : 디지털방송추진협의회 TV분과위원
주관심분야 : 디지털방송, 영상부호화, 멀티미디어전송

진 경 식

2000년 2월 : 군산대학교 전파공학과 공학사
2000년 3월 : 군산대학교 전파공학과 석사과정
주관심분야: 디지털 방송, 영상부호화

한 학 수

2000년 2월 : 군산대학교 전파공학과 공학사
2000년 3월 : 군산대학교 전파공학과 석사과정
주관심분야: 디지털 방송, 영상부호화

최 준 영

1997년 2월 : 성균관대학교 전자공학과(공학사)
1999년 2월 : 광주과학기술원 정보통신공학과(공학석사)
1999년 3월 ~ 현재 : 한국전자통신연구원(ETRI) 방송시스템연구부 근무
주관심분야 : 디지털 방송 시스템, EPG 응용 기술, 데이터 방송, 방송/통신망 연동

이 진 환

1987년 2월 : 한국항공대학교 통신공학과 졸업
2000년 2월 : 한국정보통신대학원 통신공학부 석사학위 취득
1989년 2월 ~ 현재 : 한국전자통신연구원(ETRI) 방송시스템연구부 AV전송연구팀 팀장