論文96-33A-7-3

줄길이 신호원의 순환지수 부호화

(Encoding of a Run-Length Source Using Recursive Indexing)

徐 在 俊 *, 羅 相 臣 **

(Jaejoon Seo and Sangsin Na)

요 약

본 논문에서는 팩시밀리 신호원을 위한 순환지수 이진부호를 설계하고 그 성능을 고찰하였다. 여기에 사용된 신호원은 고해상도 팩시밀리 신호의 백색 화소 줄길이이며, 성능의 비교 대상으로는 G.3 팩시밀리에 사용되는 수정 허프만 부호를 설정하였다. 순화지수화는 기하분포 신호원의 엔트로피를 보존하고 있다고 알려져 있는데, 기하분포에 유사한 실제 신호원을 순환지수화한 경우 신호원 엔트로피의 2% 오차 이내로 보존됨을 관찰하였다. 순환지수 이진부호는 순환지수화를 거친 신호원에 허프만 알고리즘을 적용하여 설계되었다. 이렇게 설계된 부호는 거의 문자만으로 이루어진 문서형 신호원과 표, 그래프, 또는 그림들로 구성되어 있는 그림형 신호원에 각각 적용되고, 부호화의 효율을 수정 허프만 부호의 효율과 비교 분석하였다. 수정 허프만 부호는 문서형 신호원에 대해서는 그 효율이 좋으나, 그림형 신호원에 대해서는 효율이 좋지 않다는 결과를 얻었다. 이에 비하여 순환지수 이진부호는 기하 분포와 유사한 분포를 하는 그림형 신호원에 적용할 때, G.3 팩시밀리의 수정 허프만 부호보다 그 요율이 신호원 엔트로피의 8%에서 20%까지 작아져 상당히 개선된 성능을 보임을 관찰하였다.

Abstract

This paper deals with the design of a recursively-indexed binary code for facsimile sources and its performance. Sources used here are run-lengths of white pixels from higher-resolution facsimile. The modified Huffman code used for G.3 facsimile is chosen for the performance comparison. Experiments confirm the fact that recursive indexing preserves the entropy of a memoryless geometric source: the entropy of a recursively-indexed physical source with roughly geometric distribution remains within 2% of the empirical source entropy. The designed recursively-indexed binary codes consist of a recursive indexing and an ensuing Huffman code. The performance of a recursively-indexed binary code applied to text-type documents and to graphics-type documents is compared with that of the modified Huffman code. Numerical results show that the modified Huffman code performs well for text-type documents and not equally well for graphics-type documents. On the other hand, recursively-indexed binary codes have shown a better performance for graphics-type documents whose distribution are similar to a geometric distribution. Specifically, the code rates of recursively-indexed binary codes with 60 codewords are from 8% to 20% of the empirical source entropy smaller than that of the modified Huffman code with 91 codewords.

Ⅰ. 서 론

순환지수 이진부호는 그림 1과 같이 순환지수법과 이진부호의 두 단계로 구성된다. 신호원 기호들로 이루어진 신호열 $S_1S_2S_3$ …은 순환지수화를 통하여 표시 기호열들로 이루어진 기호열 $I_1I_2I_3$ …로 변환된다. 이는 다시 이진부호의 입력으로 사용되어 이진수열 $Z_1Z_2Z_3$ …로 바뀐다. 이 때 순환지수는 다음 단에 있는 이진부호의 입력 기호의 갯수를 조절해 주는 선행 처리기 역할을 한다고 생각할 수 있다. 이러한 선행처리기는, 특히 기하 분포와 같은 무한 갯수 기호를 갖는 신호원을 부호화할 때, 더욱 더 필요하다. 그 이유

* 學生會員, ** 正會員, 亞洲大學校 電氣電子工學部
(School of Elec. & Electronics Eng., Ajou Univ.)

接受日字: 1995年10月21日, 수정완료일: 1996年6月20日

는 보통의 이진부호화기는 부호화기의 입력은 유한 개의 기호를 갖는다고 가정하기 때문이다. 그러한 부호들은 예를 들면 잘 알려진 허프만 알고리즘이나 셰논.파노 알고리즘을 사용하여 설계된다. 순환지수법을 이용하여 무한 개의 기호를 갖는 신호를 부호화하는 방법은 [1], [2]에서 이미 연구되어졌다. 특히, 기하 분포의 신호원을 순환지수화하면 신호원의 엔트로피가 보존된다는 순환지수법의 성질에 관한 이론적인 증명을 [2]에서 제시하고 있다.

그림 1. 순환지수 이진 부호화기
Fig. 1. Recursively-indexed binary encoder.

이 논문은 무한 개의 기호를 갖는 신호원을 부호화하는데 필요한 선행처리 과정에 순환지수법을 적용하고, 그 출력을 이진 부호화하는 순환지수 이진부호화 방법을 설명하며, 이것을 기하 분포와 유사한 분포를 갖는 실제 팩시밀리 신호원에 적용하여 [2]에서 이론적으로 증명된 순환지수법의 성질을 고찰한다.

또, 고정 이진부호어를 갖는 순환지수 부호를 다양한 신호원에 적용할 때의 효율을, CCITT 권고안의 G.3 팩시밀리에 사용되는 수정 허프만 부호의 효율과 비교한다[3]. 이때 순환지수 이진부호화에 사용된 고정 이진부호어는 다음 네 종류를 사용한다. 첫째, 그림, 표, 테이블을 포함하는 그림형 신호원들의 전체적인 분포를 이용하여 설계된 것, 둘째, 문자들만으로 구성된 문서형 신호원들의 전체적인 분포를 이용하여 설계된 것, 세째, 그림형과 문서형을 혼합한 전체적인 분포를 이용하여 설계된 것, 네째, 기하 분포와 유사한 분포를 갖는 신호원들의 전체적인 분포를 이용하여 설계된 것들이다.

본 논문에서는, 기하 분포와 유사한 신호원을 순환지수 이진부호화하였을 때, 문서형 고정부호어를 제외한 다른 종류의 고정부호어를 사용한 경우 수정 허프만 부호보다 부호 요율이 신호원 엔트로피의 8%에서 20%까지 감소되는 개선된 성능의 결과를 얻었다.

이것은 특정 신호원에 최적화된 부호어가 아닌 일반적인 고정부호어를 설계하여 적용한 결과이고, 또 수정 허프만 부호에 사용되는 고정부호어 갯수의 $\frac{2}{3}$ 수준에 해당하는 고정부호어를 사용했을 때의 결과라는 점에서 본 논문의 의미를 찾을 수 있다. 즉, 기억 용량, 처리 속도 등과 관계되는 시스템적인 복잡도를 줄이면서 성능이 개선된 부호를 얻은 것이다.

이 논문은 제 II 장에서 순환지수법에 대해 설명하고, 제 III 장에서는 설명된 순환지수 이진부호를 기하 분포와 유사한 팩시밀리 신호원에 적용할 때의 실험적 엔트로피의 변화를 고찰한다. 제 IV 장에서는 앞서 밝힌 네 종류의 고정 이진부호어를 설계하여 사용할 때, 각각의 경우에 대한 순환지수 이진부호의 효율을 CCITT 권고안의 수정 허프만 부호의 효율과 비교 분석한다.

II. 순환지수 부호화

● 대응 관계

그림 1의 순환지수 이진부호에서 이진부호앞에 사용되는 순환지수는 신호원의 기호(신호원 기호 집합)를 유한 갯수를 갖는 표시기호 집합 원소의 열(string)로 표현하는 대응 관계로 취급될 수 있다. 예를 들면 무한히 많은 원소들로 이루어진 신호원 기호 집합(alphabet)을 A, 표시기호 집합(representation set)을 B라 하고, 편의상 $A=\{0,1,2,3,4,5,6,\cdots\}$, 그리고 $B=\{0,1,2,3,\cdots M-1\}$이라 하고, 이 두 집합은 올림순으로 정렬되어있다고 하자. 집합 B에 의한 A의 순환지수 I는 다음과 같이 정의한다[2]. A의 원소 I가 만일 $i=q(M-1)+r$이면,

$$I(i) = \underbrace{(M-1)(M-1)\cdots(M-1)}_{q\text{ 번}},r \qquad (1)$$

여기서 $(M-1)\cdots(M-1)r$은 기호들의 접속된 열(concatenated string)이다.

이때 q는 몫, $M-1$은 B의 정렬된 원소들 중 마지막 원소, r은 나머지를 나타낸다.

다음 [예 1]은 무한 개의 신호원 기호 집합 A를 유한 개의 표시기호 집합 B를 이용하여 신호열을 순환지수화하고 복호화하는 예이다.

[예 1] 순환지수화와 복호화

$A=\{0,1,2,3,4,5,6,\cdots\}$이고, $B=\{0,1,2,3\}$이라 하자. 그러면 $M=4$이다. 이때 순환지수법은 표 1과 같은 대응 관계로 나타낼 수 있다. 예를 들어 A의 원소 '6'이 순환지수열 '330'으로 대응되는 과정은 $6 = 2 \times 3$

+0 이므로 $q=2$, $r=0$이다.

따라서 정의식 1에 의하여 $I(6)$은 기호열 '330'이다. 순환지수화와 복호화는 다음과 같이 설명될 수 있다. 예를 들어 신호열이 $S_1 S_2 S_3 \cdots S_{14}$ = |0|1|3|2|0|6|5|4|0|2|1|0|1|3|일 경우에 순환지수화하면, 이에 대응하는 순환지수열로 구성된 기호열은 $I_1 I_2 I_3 \cdots I_{14}$ = |0|1|30|2|0|330|32|31|0|2|1|0|1|30|이다.

여기서 '|' 표는 단지 부호어를 구별하기 위해 사용되었으며, 수신측에 전송되지 않는다. 기호열 $I_1 I_2 I_3 \cdots I_{14}$는 다시 이진부호화되어 $Z_1 Z_2 Z_3 \cdots$로 변환되고 왜곡이 없는 전송로를 거쳐 수신측에 전송된다. 수신측에서는 수신된 신호 $Z_1 Z_2 Z_3 \cdots$를 이진 복호를 거쳐 기호열 $I_1 I_2 I_3 \cdots$ = 01302033032310210130을 얻는다. 순환지수 복호는 표 1의 역 관계이므로 다음의 과정을 통하여 복호화할 수 있다.

(a) 수신 기호가 최초로 '3'이 아닐 때까지 읽는다.
(b) 읽힌 기호에 해당하는 수를 모두 더한다.

표 1. 신호원 기호 집합 A 와 순환 지수열 $I(i)$의 대응 관계
Table 1. Relation of source symbol set A and recursively-indexed string $I(i)$.

신호원 기호 집합 A	순환 지수열 $I(i)$
0	0
1	1
2	2
3	30
4	31
5	32
6	330
⋮	⋮

이러한 복호화 과정을 적용하면, 위의 기호열 중 첫 번째 기호 '0'은 '3'이 아니므로, 이때의 '0'은 한 개의 신호원 기호로 복호된다. 두 번째 기호 '1' 또한 신호원 기호 '1'로 복호된다. 세 번째 기호는 '3'이므로 그 다음 기호를 읽는다. 네 번째 기호는 '3'이 아닌 '0'이므로 '30'이 한 개의 신호원 기호를 표시하고, 이것은 3+0=3에 의해 신호원 기호 '3'으로 복호된다. 이 방법으로 위의 기호열을 복호화하면 |01320654021013|을 얻으며, 이 열은 원래의 신호원 열과 일치한다. □

위의 예로부터 순환지수법은 입력 고정-출력 가변 길이인 첨두없는 부호(prefix-free code)로서 순시 복호성이 있으며 유일 복호성이 있음을 확인할 수 있다.

● **표시기호의 통계적 분포**

순환지수 부호는 표시기호 B의 분포에 최적인 입력 고정-출력 가변 길이 이진부호를 사용하면, 최적의 순환지수 이진부호를 얻을 수 있다. 최적의 부호 설계를 위해서는 표시기호의 통계적 분포를 알아야 되며, 이 분포는 원신호 A의 분포로부터 유도할 수 있다. 그림 1의 신호원 $S_1 S_2 S_3 \cdots$에서 $p_i = \Pr(S_k = i)$라 하면 모든 i에 대해 $p_i \geq p_{i+1}$가 되도록 순서가 정렬되어 있다고 가정하자. 신호열이 주어져 있을 때, 이로부터 표시기호의 발생 수와 이를 이용한 상대 빈도수를 구하는 과정을 [예 2]에 나타내었다.

[예 2] 표시기호 B의 분포

신호원의 열이 주어진 경우에, 순환지수화된 신호열의 분포를 구해 보기로 한다. [예 1]에 사용된 신호원과 표시기호 집합을 이용하여 표시기호의 통계적 분포를 구해 보자. 먼저 표시기호 '0'의 발생 수 n_0는 다음과 같이 구한다. 신호원 기호 '4', '3', '6'이 나타날 때, 표시기호 '0'은 한 번씩 나타나며 신호원 기호 '0'의 상대 빈도는 신호원 기호 '0'의 반복되는 수를 신호원 기호들의 반복되는 수의 총합으로 나누어 구한다. [예 1]의 원신호열에 14개의 신호원 기호들이 있고 그 중 '0'은 '4'번, '3'은 '2'번, 그리고 '6'은 1번씩 각각 나타난다. 따라서, 표시 기호 '0'의 발생 수 n_0는, 신호원 기호 i의 발생 수를 n_{Si}라 할 때, $n_0 = n_{S0} + n_{S3} + n_{S6} = 4+2+1=7$이다. 이와 유사한 방법에 의해 표시기호의 발생 수는 $n_1 = n_{S1} + n_{S4} = 3+1=4$, $n_{2=n_{S2}} + n_{S5} = 2+1=3$이다. 표시기호들 중 가장 큰 수인 '3'의 발생 수는, 다른 표시기호들과는 다르게 한 신호원 기호에 여러 번 반복되어질 수 있는 점을 고려해야 한다. 예를 들면, 신호원 기호 '6'은 2개의 표시기호 '3'과 1개의 표시기호 '0'으로 구성되어 있으므로, 표시기호 '3'의 발생 수는 $n_3 = n_{S3} + n_{S4} + n_{S5} + 2\times n_{S6} = 2+1+1+2\times 1 = 6$ 이다.

표시 기호의 상대 빈도수 q_j 는 다음과 같이 계산된다.

$$q_j = \frac{n_j}{\sum_{i=0}^{3} n_i}, \quad j=0,1,2,3 \tag{2}$$

여기서 n_j는 표시기호 'j'의 발생 빈도수이다. 따라서, $q_0=\frac{7}{20}, q_1=\frac{4}{20}, q_2=\frac{3}{20}, q_3=\frac{6}{20}$ 을 얻는다. 이 q_j는 순환지수법에 의해 신호원으로부터 유도되는 표시기호의 확률 분포로 간주될 수 있다. □

● 확장인자

순환지수법은, 무한 갯수의 신호원 기호들로 구성된 신호열을 유한 개의 표시기호로 표시하기 위해서, 표시기호를 반복적으로 사용한다. 표시기호의 반복 사용에 의해서 하나의 신호원 기호가 표시되는 표시 기호의 평균 갯수를 순환지수법 I의 확장 인자라하고, 이를 ϵ으로 표시하기로 하자[2]. 위의 [예 2]에 사용된 신호열에 대한 확장 인자 ϵ의 값을 구하면,

$$\epsilon=\frac{\sum_{j=0}^{3} n_j}{\sum_{i=0}^{\infty} n_{Si}}=\frac{7+4+3+6}{14}=1.429 \qquad (3)$$

여기서 $\sum_{j=0}^{3} n_j$는 표시기호열에 나타난 표시기호들의 전체 수, $\sum_{i=0}^{\infty} n_{Si}$는 신호열에 나타난 신호원 기호의 전체 수를 각각 나타낸다. 따라서, 신호원 기호 1개를 표시하는데 약 1.43개의 표시기호가 필요하다. 확장 인자 ϵ는 신호원 기호와 순환지수화를 통해 재구성된 이진부호화기의 입력 기호와의 관계를 나타내고, 표시기호의 갯수가 매우 커지면 이 인자는 1에 근접한다.

● 순환지수법의 성질

순환지수화기 다음 단의 입력 고정-출력 가변 길이 이진부호화기는 한 번에 한 개의 표시기호를 받아서 이에 해당하는 가변 길이 부호어를 출력한다.

표시기호를 허프만 부호화할 때에, 표시기호에 대한 평균 이진부호의 갯수인 표시기호 부호 요율(code rate) R_i는 다음의 범위에 있다.

$$H(B)\le R_i < H(B)+1 \qquad (4)$$

여기서 $H(B)$는 순환지수화에 의한 표시기호의 엔트로피이며, $H(B)=-\sum_{j=0}^{M-1} q_j \log_2 q_j$ 이다.

여기서, q_j는 표시기호 j의 상대 빈도수이다. 순환지수화한 다음, 최적의 입력 고정-출력 가변 길이 이진부호화한 신호원 부호 요율 R_S(즉 신호열 $S_1 S_2 S_3 \cdots$의 한 기호를 표시하는 데 필요한 신호원 기호당 평균 이진부호의 갯수)는, 따라서 ϵR_i가 되고, 다음의 범위 안에 있게 된다.

$$\epsilon H(B)\le R_S < \epsilon(H(B)+1) \qquad (5)$$

순환지수법은 일대일 대응 관계이나, 일반적으로 신호원 엔트로피 $H(S)\neq\epsilon H(B)$이다. 이는 신호원 기호들이 표시기호열로 표시되고, $H(B)$를 계산할 때 표시**기호열**의 확률이 아니라 표시**기호**의 확률이 사용되기 때문이다. 그러나, 기하 분포 확률 질량 함수 $p_i=\Pr(S=i)=(1-p)p^i$, $i=0,1,2,\cdots$를 갖는 독립 동일 분포된(independent identically-distributed) 신호원 S는 순환지수 부호화했을 때 $H(S)=\epsilon H(B)$가 성립하여 엔트로피가 보존된다[2].

Ⅲ. 순환지수법을 이용한 기하 분포의 부호화의 실험적 고찰

기하 분포 신호원을 순환지수 이진부호화하면 신호원의 엔트로피는 보존되지만 실제 신호원은 정확한 기하 분포를 하지 않는다. 따라서 기하 분포와 유사한 실제 신호원을 순환지수 이진부호화할 때, 신호의 실험적 엔트로피 $H_{emp}(S)$와 순환지수화 후의 엔트로피 $\epsilon H_{emp}(B)$와의 관계를 고찰하고자 한다. 여기서 $H(B)$라하지 않고 $H_{emp}(B)$라 한 이유는 실제 신호원으로부터 확률 분포를 추정하여 실험적으로 구한 엔트로피이기 때문이다. 이 장에서는 $H_{emp}(S)$, $\epsilon H_{emp}(B)$, 그리고 표시기호의 갯수가 $\epsilon H_{emp}(B)$의 변화에 미치는 영향에 대해 고찰하고자 한다.

이 논문에서는 그림 2와 같은 팩시밀리 신호원의 연속되는 백색 화소 줄길이를 신호원 기호로 설정한다. 신호원의 전체 화소수는 가로×세로가 1,728 × 2,576 =4,105,728개이고, 표준 A4 크기의 고해상도 팩시밀리의 신호원에 해당한다[3]. 신호원의 백색 화소 줄길이에 의한 신호원 기호 집합을 A_W, 흑색은 A_B라 했을 때 다음의 결과가 관찰된다.

$$A_W=\{1,2,\cdots,1728\}, A_B=\{1,2,\cdots,110\} \qquad (6)$$

이들 흑색 화소의 줄길이는 백색에 비해 현저히 작은데, 이는 대부분의 문서는 백색 바탕에 흑색 문자, 표, 그림으로 구성되어 있기 때문이다.

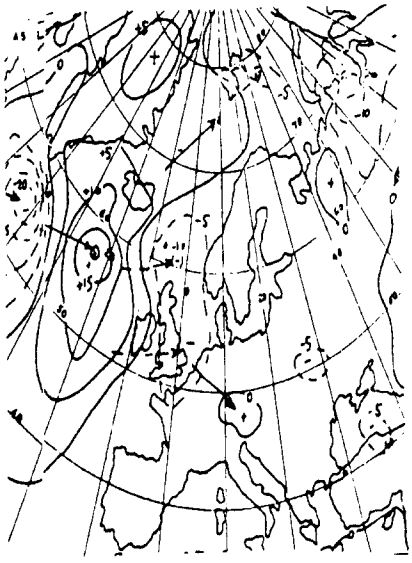


그림 2. 기상도
Fig. 2. The weathermap.

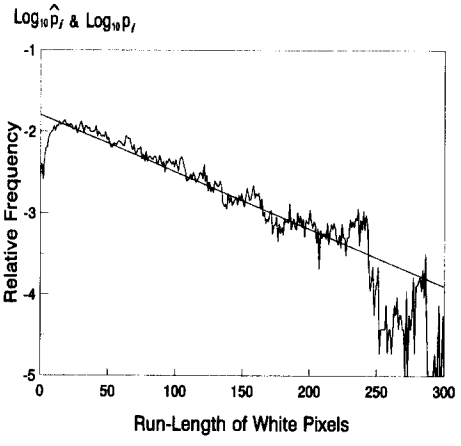


그림 3. 그림 2 기상도의 백색 화소의 상대적 빈도 분포
Fig. 3. The distribution of the relativ efrequency of run-length of white pixels of the weathermap of Fig. 2.

이 때 백색 화소 줄길이에 의한 신호원 기호의 상대 빈도수를 구하면 다음과 같다.

$$\hat{p_i} = \Pr(S=i) = \frac{n_{Si}}{\sum_{i=1}^{1728} n_{Si}} \tag{7}$$

여기서 S는 신호, n_{Si}는 신호원 줄길이 i의 빈도이다. 그림 3은 그림 2의 백색 화소 줄길이의 상대 빈도수 $\hat{p_i}$의 분포와 $p=0.984$인 기하 분포, $p_i=(1-p)p^i$를 보였다. 그림 3에서 신호원 기호 i를 300까지 나타내었는데, 이는 신호원 기호 1에서 300까지의 분포가 전체 분포의 99%에 해당하기 때문이다. 이 그림으로 부터 백색 화소 줄길이의 분포는 기하 분포와 유사함을 관찰할 수 있다. 이 분포를 갖는 백색 화소에 의한 신호원의 실험적 엔트로피는 $H_{emp}(S) = -\sum_{i=1}^{1728} \hat{p_i} \log_2 \hat{p_i}$에 의해 $H_{emp}(S) = 7.415$ bits/symbol이다. 또한 $p=0.984$인 기하 분포의 엔트로피는 7.397 bits/symbol로 엔트로피 측면에서도 이 신호원은 기하 분포에 근접함을 볼 수 있다.

백색 화소 줄길이에 의한 신호원 기호 집합 A_W의 원소들로 이루어진 신호원 $S_1 S_2 S_3 \cdots$는 제 II 장에서 설명한 과정에 의해 순환지수화된다. 순환지수의 출력 $I_1 I_2 I_3 \cdots$은 표시기호만으로 구성되어 있고, 이 때의 표시기호들의 상대 빈도수를 구하고 $H_{emp}(B)$를 구한다. 또는 [예 2]에서처럼 신호원의 통계 분포를 이용하여 표시기호들의 상대 빈도수를 구할 수 있다.

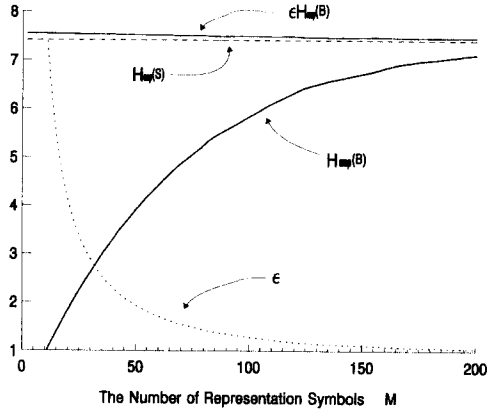


그림 4. $H_{emp}(S)$와 $\varepsilon H_{emp}(B)$의 비교
Fig. 4. Comparison $H_{emp}(S)$ and $\varepsilon H_{emp}(B)$.

그림 4는 표시기호 갯수의 변화에 따른 그림 2의 백색 화소 줄길이의 실험적 엔트로피 $H_{emp}(S)$, 확장 인자 ε, 표시기호의 실험적 엔트로피 $H_{emp}(B)$, 그리고 $\varepsilon H_{emp}(B)$의 관계를 나타낸다. 표시기호 갯수가 증가할수록 은 작아지고, $H_{emp}(B)$는 커진다. ε 그림으로부터 표시기호 갯수의 증가에 따른 ε의 감소 변화율과 $H_{emp}(B)$의 증가 변화율은 비슷하고 $\varepsilon H_{emp}(B)$는 실험에서 고려된 M의 전체 범위에서 $H_{emp}(S)$의 약 2%

이내의 오차 범위 안에 있다.

또 이 값은 M=2일 때 $H_{emp}(S)$의 1.8%의 차이에서 M=200일 때 $H_{emp}(S)$의 0.8%의 차이 정도로 표시기호 갯수의 변화에 큰 영향을 받지 않음을 관찰하였다. 이 결과에 의해 제 II 장에서 기술한 것처럼 신호원 분포가 정확한 독립 동일 기하 분포를 이루고 있다면, 표시기호 갯수에 관계없이 신호원 엔트로피는 보존됨을 확인할 수 있다.

Ⅳ. 순환지수 이진부호와 팩시밀리용 수정 허프만 부호의 성능 비교

이 장에서는 여러 종류의 신호원에 대한 고정부호어를 가진 순환 지수 부호의 효율을 수정 허프만 부호와 비교한다. 이 효율은 수정 허프만 부호화한 후의 신호원 기호당 이진수의 평균 갯수, 즉 R_H와 고정 이진부호어를 갖는 순환지 수 이진부호화 후의 이진수의 평균 갯수 R_S를 각각 신호원의 엔트로피로 나눈 값 C_H와 C_S로 측정된다.

$$C_H = \frac{R_H}{H_{emp}(S)}, C_S = \frac{R_S}{H_{emp}(S)} = \frac{\varepsilon R_i}{H_{emp}(S)} \quad (8)$$

순환지수 이진부호의 성능은 고정 이진부호어에 따라 영향을 받으므로 모든 신호원들에 대해 최적화될 수 있는 일반적인 고정 이진부호어를 설계해야 한다. 이런 이진부호어를 얻기 위해서 여러 신호원들로부터 신호원 기호의 전체적인 분포를 구하여 이것을 순환지수화한 표시기호의 상대 빈도수를 구한다. 여기에 허프만 알고리즘을 적용하여 각각의 표시기호에 대응하는 부호어를 설계한다. 그리고 순환지수 이진부호는 순환지수열을 구성하는 표시기호에 이미 만든 고정 이진부호어를 대응시켜 만들어진다.

신호원은 줄길이의 장단의 정도에 따라 문서형 신호원과 그림형 신호원 두 종류로 나누었다. 그림형 신호원은 대부분 그래프, 그림 또는 표들로 구성되므로, 긴 줄길이에 의한 신호원 기호의 상대 빈도가 매우 높다. 본 논문에서는 줄길이가 100 이상인 분포가 전체의 20% 이상일 때의 신호원으로 설정한다. 문서형 신호원은 대부분 문자들로만 구성되어 있으므로, 긴 줄길이에 의한 신호원 기호의 상대 빈도는 매우 낮다. 따라서 100 이상의 줄길이 분포가 전체의 20% 이하인 신호원으로 설정하여 구분한다.

앞에서 분류한 신호원의 종류를 바탕으로 고정 이진부호어 또한 다음의 네 종류를 설계하였다.

(1) 문서형 고정부호어

문서형을 대표할 수 있는 신호원 즉, 그림 5(a)와 비슷한 형태의 신호원들의 백색 화소 줄 길이에 의한 신호원 기호의 전체 분포를 구하고, 순환지수화한 후, 허프만 알고리즘을 적용한 이진부호화하여 고정부호어를 설계하였다.

(2) 그림형 고정부호어

그림형을 대표할 수 있는 신호원, 그림 5(b) 또는 (c)와 비슷한 형태의 신호원들을 이용하여 (1)의 문서형 고정부호어를 만드는 과정과 같은 방법으로 고정부호어를 만든다.

(3) 혼합형 고정부호어

그림형과 문서형 신호원들의 혼합된 전체 분포를 이용하여 위와 같은 방법으로 고정부호어를 만든다.

(4) 기하 분포형 고정부호어

신호원들중 기하 분포와 유사한 분포를 갖는 신호원, 그림 5(d)와 비슷한 형태의 신호원들의 전체 분포를 이용하여 고정부호어를 만든다.

이러한 네 종류의 고정부호어를 이진부호로 사용할 때, 여러 종류의 신호원들에 대해 순환지수 이진부호의 효율을 수정 허프만 부호의 효율과 비교한다. 이 비교에 사용된 신호원들은 CCITT 시험문서로서 문서형 신호인 시험문서 4 (그림 5(a) 신호원 1), 그림형 신호인 시험문서 2 (그림 5(b) 신호원 2), 시험문서 8 (그림 5(c) 신호원 3), 그리고 기하 분포와 유사한 분포를 하고 있는 우리 나라 기상도인 그림 5(d) 신호원 4이다. 고정부호어를 설계할 때 사용된 신호원들은 그림 5의 신호원들을 포함하지 않는다. 즉, 앞에서 만든 고정부호어는 그림 5의 신호원들에 특별히 최적화된 부호어가 아닌 일반적인 고정부호어라는 점을 상기할 필요가 있다.

표 2는 C_H, 표시기호 갯수 $M=60$일 때 C_S, 그리고 P_{HS}의 값을 보여준다. 여기서 P_{HS}는 R_H와 R_S의 차이를 $H_{emp}(S)$에 대한 백분율로 나타낸 값으로 다음 식과 같다.

L'ordre de lancement et de réalisation des applications fait l'objet de décisions au plus haut niveau de la Direction Générale des Télécommunications. Il n'est certes pas question de construire ce système intégré "en bloc" mais bien au contraire de procéder par étapes, par paliers successifs. Certaines applications, dont la rentabilité ne pourra être assurée, ne seront pas entreprises. Actuellement, sur trente applications qui ont pu être globalement définies, six en sont au stade de l'exploitation, six autres se sont vu donner la priorité pour leur réalisation.

Chaque application est confiée à un "chef de projet", responsable successivement de sa conception, de son analyse-programmation et de sa mise en oeuvre dans une région-pilote. La généralisation ultérieure de l'application réalisée dans cette région-pilote dépend des résultats obtenus et fait l'objet d'une décision de la Direction Générale. Néanmoins, le chef de projet doit dès le départ considérer que son activité a une vocation nationale donc refuser tout particularisme régional. Il est aidé d'une équipe d'analystes-programmeurs et entouré d'un "groupe de conception" chargé de rédiger le document de "définition des objectifs globaux" puis le "cahier des charges" de l'application, qui sont adressés pour avis à tous les services utilisateurs potentiels et aux chefs de projet des autres applications. Le groupe de conception comprend 8 à 10 personnes représentant les services les plus divers concernés par le projet,et comporte obligatoirement un bon analyste attaché à l'application.

II - L'IMPLANTATION GEOGRAPHIQUE D'UN RESEAU INFORMATIQUE PERFORMANT

L'organisation de l'entreprise française des télécommunications repose sur l'existence de 20 régions. Des calculateurs ont été implantés dans le passé au moins dans toutes les plus importantes. On trouve ainsi des machines Bull Gamma 30 à Lyon et Marseille, des GE 425 à Lille, Bordeaux, Toulouse et Montpellier, un GE 437 à Massy, enfin quelques machines Bull 300 TI à programmes câblés étalent récemment ou sont encore en service dans les régions de Nancy, Nantes, Limoges, Poitiers et Rouen ; ce parc est essentiellement utilisé pour la comptabilité téléphonique.

A l'avenir, et la plupart des fichiers nécessaires aux applications décrites plus haut peuvent être gérés en temps différé, un certain nombre d'entre eux devront nécessairement être accessibles, voire mis à jour en temps réel : parmi ces derniers le fichier commercial des abonnés, le fichier des renseignements, le fichier des circuits, le fichier technique des abonnés contiendront des quantités considérables d'informations.

Le volume total de caractères à gérer en phase finale sur un ordinateur ayant en charge quelques 500 000 abonnés a été estimé à un milliard de caractères au moins. Au moins le tiers des données seront concernées par des traitements en temps réel.

Aucun des calculateurs énumérés plus haut ne permettait d'envisager de tels traitements.

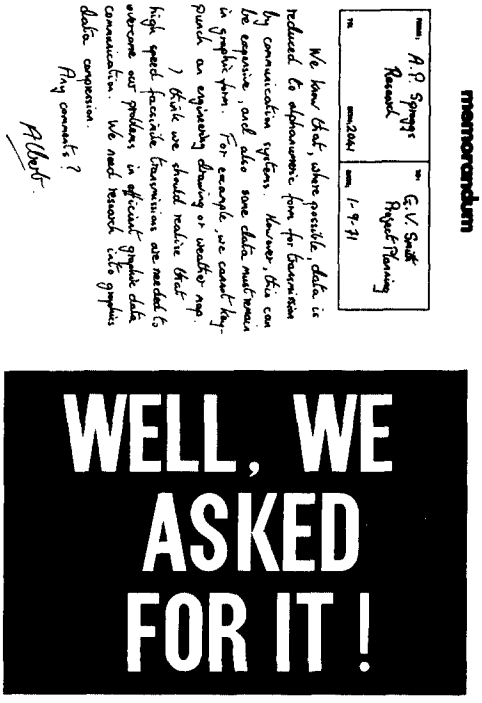
L'intégration progressive de toutes les applications suppose la création d'un support commun pour toutes les informations, une véritable "Banque de données", répartie sur des moyens de traitement nationaux et régionaux, et qui devra rester alimentée mise à jour en permanence, à partir de la base de l'entreprise, c'est-à-dire les chantiers, les magasins, les guichets des services d'abonnement, les services de personnel etc.

L'étude des différents fichiers à constituer a donc permis de définir les principales caractéristiques du réseau d'ordinateurs nouveaux à mettre en place pour aborder la réalisation du système informatif. L'obligation de faire appel à des ordinateurs de troisième génération, très puissants et dotés de volumineuses mémoires de masse, a conduit à en réduire substantiellement le nombre.

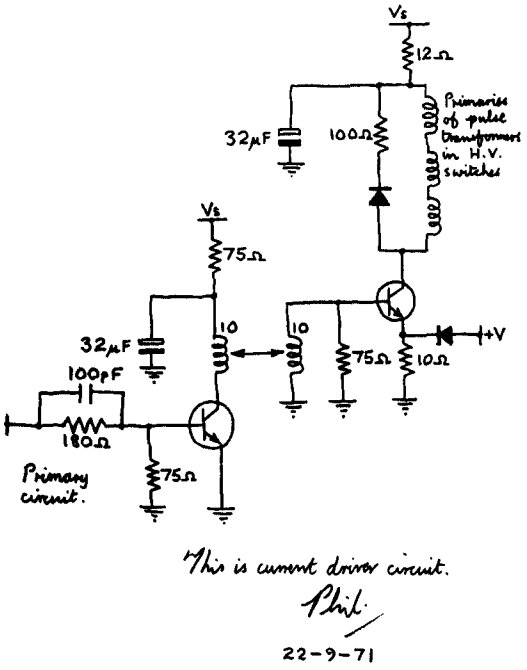
L'implantation de sept centres de calcul interrégionaux constituera un compromis entre : d'une part le désir de réduire le coût économique de l'ensemble, de faciliter la coordination des équipes d'informaticiens, et d'autre part le refus de créer des centres trop importants difficiles à gérer et à diriger,et posant des problèmes délicats de sécurité. Le regroupement des traitements relatifs à plusieurs régions sur chacun de ces sept centres permettra de leur donner une taille relativement homogène. Chaque centre "gèrera" environ un million d'abonnés à la fin du VIème Plan.

La mise en place de ces centres a débuté au début de l'année 1971 : un ordinateur IRIS 50 de la Compagnie Internationale pour l'Informatique a été installé à Toulouse en février ; la même machine vient d'être mise en service au centre de calcul interrégional de Bordeaux.

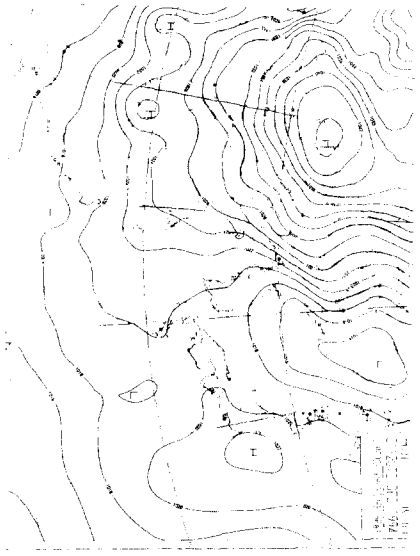
(a)



(c)



(b)



(d)

그림 5. 팩시밀리 신호원
(a)신호원 1 (b) 신호원 2 (c) 신호원 3 (d) 신호원 4

Fig. 5. Fascimile sources.
(a) Source 1 (b) Source 2 (c) Source 3 (d) Source 4

$$P_{HS}=\frac{(R_H-R_S)}{H_{emp}(S)}\times 100\% \quad (9)$$

따라서, 이 표에서 P_{HS}값이 양수이면서 클 수록, 수정 허프만 부호보다 더 좋은 성능을 나타낸다. 여기서 C_H와 비교되는 C_S값을 표시기호 갯수 $M=60$인 경우로 설정한 것은 전체적인 결과로 볼 때 $M=60$ 정도 될 때 표시기호 갯수에 따른 부호 요율의 변화가 안정되기 때문이다. 참고로 G.3 팩시밀리에 사용하는 수정 허프만 부호는 줄 끝점(End Of Line) 부호어를 제외하고 91개의 부호어가 있으므로 60개의 표시기호의 갯수, 즉 고정부호어의 수는 이것의 $\frac{2}{3}$ 수준이다.

그림 6은 표 2의 값을 신호원에 따라 각각 네 종류의 고정 이진부호어를 사용할 때의 C_H, C_S를 나타낸다. 무손실 신호원 부호화 정리에 의하면 신호원 부호당 평균 이진부호수, 즉 R_S는 항상 $H(S)$보다 크거나 같아야 한다. 그러므로, 그림 6의 값들은 1보다 작아질 수 없다. 그리고 그 값이 1에 근접하는 것은 R_S가 $H(S)$에 근접함을 의미하고, 이것은 부호화의 효율이 우수함을 나타낸다. 그림 5의 네 개의 신호원에 대한 수정 허프만 부호의 효율은 문서형 신호원(신호원 1)에서 효율이 가장 좋고, 그림형 신호원에서는 효율이 떨어진다. 신호원 1(문서형 신호원)을 60개의 고정 이진부호어를 가진 순환지수 부호화하면 문서형에 적합한 고정 이진부호어를 사용했을 때 효율이 가장 좋지만, 수정 허프만 부호보다 R_S가 $H_{emp}(S)$의 13% 정도 증가하여 효율이 나쁘다. 또 신호원 2(그림형 신호원)는 문서형에 적합한 고정부호어를 사용할 때 순환지수 이진부호의 효율이 가장 떨어지고, 기하 분포에 적합한 고정부호어를 사용할 때와 그림형에 적합한 고정부호어를 사용할 때 효율이 좋지만, 수정 허프만 부호보다 $H_{emp}(S)$의 14% 정도 요율이 크므로 성능이 나쁘다. 그러나, 동일한 그림형 신호원일지라도 신호원 3은 기하 분포형에 적합한 고정부호어를 사용했을 때 수정 허프만 부호의 요율보다 $H_{emp}(S)$의 약 20% 정도 감소되는 현저한 성능 개선을 보인다. 신호원 4를 60개의 고정부호어를 갖는 순환지수 이진부호화 하면, 기하 분포형에 적합한 고정부호어를 사용했을 때 R_H보다 R_S는 신호원 엔트로피의 16% 정도 작아진 가장 적은 신호원 부호 요율을 얻어, 가장 개선된 성능을 얻었다. 전체적으로 신호원 4에 대하여 사용된 고정부호어 종류에 따른 순환지수 이진부호화의 결과는, 문서형에 적합한 고정부호어를 사용한 경우를 제외하고 신호원 부호 요율이 수정 허프만 부호보다, $H_{emp}(S)$의 9%에서 16%까지 작아지는 성능 개선을 보였다. 또한 이 실험에서 사용된 어떤 고정부호어보다 기하 분포형에 적합한 고정부호어를 사용했을 때 신호원 부호 요율이 가장 $H_{emp}(S)$에 근접한다.

그림 7에서 10까지는 그림 5를 표시 기호의 갯수의 변화에 따른 고정 이진부호어를 갖는 순환지수 부호의 성능 변화를 표시한다. 수직축은 신호원 부호 요율을 신호원 엔트로피로 나눈 값의 상용로그 값이다. 따라서 수직축의 값 '0'은 $H_{emp}(S)$를 의미한다. 이미 설명한 바와같이 그림 7은 문서형 신호원(신호원 1)을 문서형 고정부호어로 순환지수 이진부호화할 때 효율이 가장 좋으나, 수정 허프만 부호화의 경우에 비하여 효율이 낮다. 문서형 고정부호어를 사용한 경우 표시기호 갯수가 60, 200일 때 R_S는 수정 허프만 부호화한 경우보다 각각 $H_{emp}(S)$의 13%, 0.4% 작은 정도로 그 차이가 감소하여 표시 기호 갯수가 증가할 수록 효율이 더욱 좋아진다. 그러나, 고정부호어의 증가는 실제로 구현할 때의 복잡성, 요구되는 기억 용량의 증가, 그리고 처리 속도의 감소로 인하여 큰 의미가 없다.

표 2. 고정 이진 부호를 갖는 순환 지수 부호 $M=60$와 수정 허프만 부호의 효율 비교
Table 2. The performance comparison of the recursively-indexed binary codes with $M=60$.

신호원	$H_{emp}(S)$	C_H	문서형부호어		기하분포형 부호어		그림형 부호어		혼합형 부호어	
			C_S	P_{HS}(%)	C_S	P_{HS}(%)	C_S	P_{HS}(%)	C_S	P_{HS}(%)
1	4.6978	1.1121	1.2431	-12.99	1.4851	-37.30	1.4598	-34.76	1.2637	-15.16
2	8.1632	1.1912	1.8323	-64.11	1.3280	-13.68	1.3443	-15.31	1.3741	-18.29
3	6.9047	1.3324	1.4069	-7.45	1.1370	19.54	1.1962	13.63	1.2564	7.60
4	7.8778	1.1879	1.2350	4.70	1.0271	16.07	1.0631	12.66	1.0992	8.86

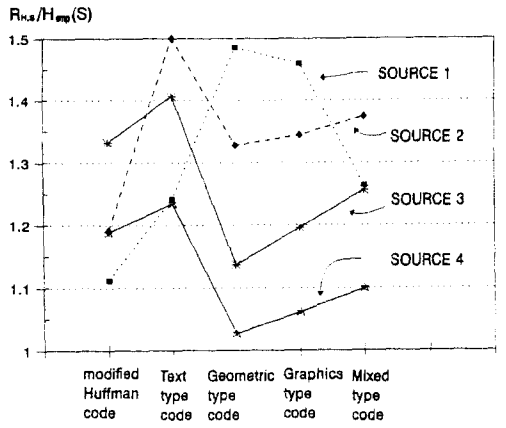


그림 6. $M=60$인 순환지수 부호와 수정 허프만 부호의 성능 비교

Fig. 6. The performance comparison of recursively-indexed binary code with $M=60$ and the modified Huffman code.

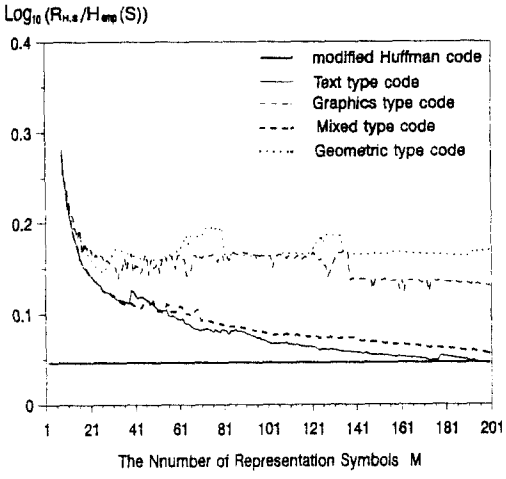


그림 7. 신호원 1에 대한 여러 가지 부호의 효율

Fig. 7. The performances of various codes for Source 1.

신호원을 순환지수화한 다음 표시기호의 통계적 분포를 구하고, 허프만 알고리즘을 적용하여 εR_i를 구하는 과정을 표시기호 갯수를 변화하면서 실험한 결과, εR_i는 표시기호의 갯수가 증가함에 따라 거의 단조 감소함을 관찰했다. 즉, 순환지수화기 다음 단에 최적의 이진부호화기를 사용하면 εR_i는 표시기호 갯수가 증가할 수록 작은 값을 갖는다는 것을 의미한다. 따라서, 순환지수 이진부호화에 사용된 신호원에 대하여 본 논문에서 설계한 고정 이진부호어와 최적화된 이진부호어와의 차이를 실험에서 얻은 εR_i로 부터 알 수 있다.

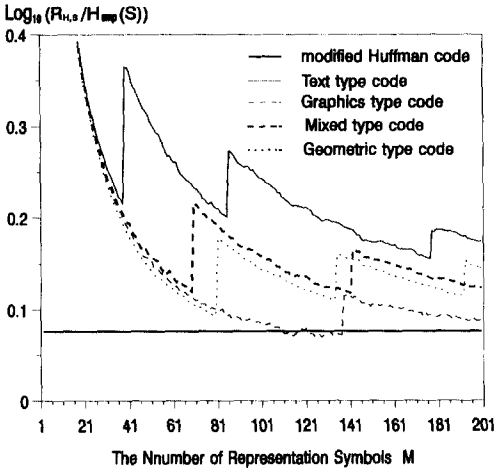


그림 8. 신호원 2에 대한 여러 가지 부호의 효율

Fig. 8. The performances of various codes for Source 2.

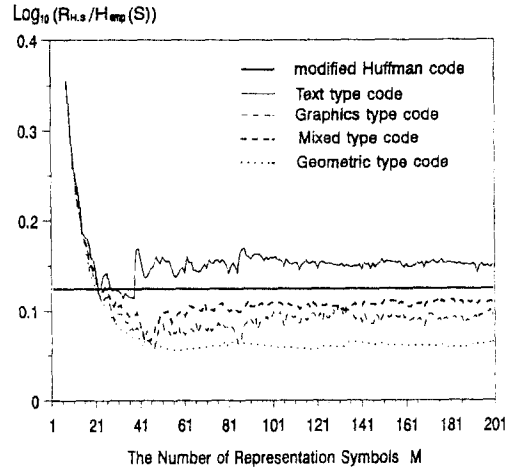


그림 9. 신호원 3에 대한 여러 가지 부호의 효율

Fig. 9. The performances of various codes for Source 3.

또한 표시기호 갯수의 변화에 따른 εR_i의 값은 부호화된 신호원이 어떤 종류인지를 구별 가능케 해주는데, 이는 신호원과 고정부호어의 형태가 같으면 다른 형태의 고정부호어를 사용할 때보다 더 작은 εR_i 를 갖기 때문이다. 예를들어 문서형 신호원을 순환지수 이진부호화할 때, 최적화된 이진 부호어는 문서형 신호원을 순환지수화한 다음, 통계적 분포를 구하고 이를 이용한

이진 부호화를 통하여 얻어진다. 이 때 최적화된 이진 부호어에 가장 근접하는 고정부호어의 종류는 문서형 고정 부호어이다. 구체적으로 말하면, 순환지수 이진 부호화에 사용된 고정부호어를 설계할 때 사용된 분포가 부호화에 사용되는 신호원의 분포에 근접할 수록 고정 부호어는 최적화된 이진 부호어에 근접한다.

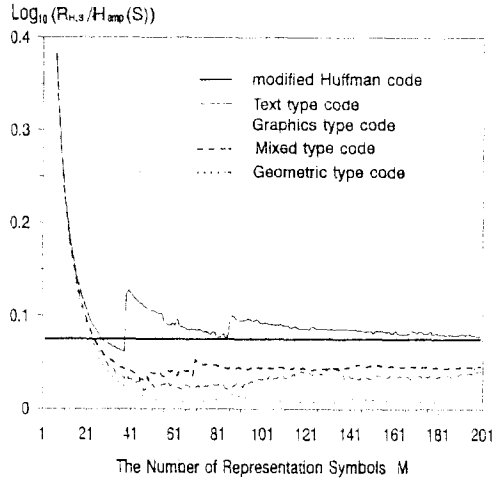


그림 10. 신호원 4에 대한 여러 가지 부호의 효율
Fig. 10. The performances of various codes for Source 4.

이러한 사실로부터 그림 7에서 그림 10까지를 분석해 보면 다음의 사실들을 고찰할 수 있다. 신호원 1은 문서형 신호원이고, 이 때 사용된 문서형 고정 부호어는 이 신호원에 최적화된 이진부호어에 근접해 있다[그림 7]. 신호원 2는 그림형 신호원이고, 사용된 그림형 고정부호어는 이 신호원에 최적화된 이진부호어와 차이가 있다[그림 8]. 신호원 3과 신호원 4는 그림형 신호원이고 또한 기하분포와 유사한 분포이다. 이 때 사용된 기하분포형 고정부호어는 최적화된 이진부호어에 근접해 있다[그림 9, 그림 10]. 그런데 신호원 4는 신호원 3보다 기하분포형 고정부호어가 이 신호원에 최적화된 이진부호어에 더 근접해 있음을 알 수 있다[그림 10]. 특히 그림 10으로부터 매우 작은 표시기호 갯수에서 εR_i가 $H_{emp}(S)$에 매우 근접하고, 표시기호 갯수의 변화에 큰 영향을 받지않음을 관찰할 수 있다.

V. 결 론

본 논문에서는 순환지수 이진부호는 기하 분포의 신호원에 대해 엔트로피가 보존된다는 성질을 이용하여, 기하 분포와 유사한 실제 팩시밀리 신호원을 순환지수 이진부호화한 다음, 표시기호 갯수를 변화시키면서 이에따른 $\varepsilon H_{emp}(B)$와 $H_{emp}(S)$를 비교 관찰하였다.

그리고 여러 종류의 신호원에 대한 수정 허프만 부호와 고정부호어를 갖는 순환지수 부호의 요율을 비교하여, 그 효율을 분석하였다. 기하 분포에 근접한 신호원에 대하여 순환지수 이진부호는 표시 기호 갯수 M에 안정되어 있어서, $\varepsilon H_{emp}(B)$는 표시 기호 갯수 $M=2, M=200$일 때 각각 신호원의 $H_{emp}(S)$의 1.8%, 0.8% 정도 오차 안에 있다. 즉, 실험한 M의 범위에서 전체적으로 2% 이내의 오차를 갖는다. 그리고, 본 논문에서 설계한 60개의 고정부호어로 구성된 순환지수 이진부호는 기하 분포와 유사한 그림형 신호원에 대해, 약 90개의 고정부호어를 갖는 수정 허프만 부호보다 신호원 부호 요율이 $H_{emp}(S)$의 8%에서 20%까지 감소되는 성능 개선을 보였다.

본 논문에서 얻은 결과는 팩시밀리 신호원의 백색 화소만을 신호원 기호로 사용한 결과이나, 백색 화소의 분포가 전체의 90%에 해당하므로 흑색 화소를 포함한 신호원을 부호화한 결과는 큰 변화가 없을 것이다.

그리고, 이진 영상에 일반적으로 많이 적용되고 있는 1차원 마코프 모델처럼 팩시밀리 신호원을 기하 분포와 유사한 분포를 갖도록 모델화하여 사용한다면 팩시밀리 성능을 더욱 향상 시킬 수 있으리라 기대한다.

참 고 문 헌

[1] K. Sayood and S. Na, "Recursively indexed quantization of memoryless sources," *IEEE Trans. on Inform. Theory*, pp. 1602--1609, Sept. 1992.

[2] S. Na, Y. K. Kim, H. S. Lee, "The Entropy of Recursively-Indexed Geometric Distribution," *Journal of Electrical Engineering and Information Science*, pp. 91-97, March, 1996.

[3] R. Hunter and A. H. Robinson, "International Digital Facsimile Coding Standards," *Proceedings of IEEE*, vol. 68, no. 7, pp. 854--867, July 1980.

[4] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, New Jersey, pp. 465--482, 1984.

저 자 소 개

徐 在 俊(學生會員)
1969년 9월 3일생. 1995년 2월 아주대학교 전자공학과 졸업(공학사). 1995년 ~ 현재 아주대학교 전기전자공학부 석사과정. 주관심 분야는 자료압축, 통신방식, 영상신호처리 등임

羅 相 臣(正會員) 第 32卷 B編 第 6號 參照
현재 아주대학교 전기전자공학부 교수