정보의 흐름으로 본 정보량의 계산

陳 庸 玉
〈경희대교수·통신공학/본지편집고문〉

1. 데이터 정보량의 계산법

이번호에서는 '88년 3월호에 기술한 정보이동에 관하여 단계적으로 구분해 보기로 한다. 정보는 7단계를 거쳐 이동을 하며 이때마다 정보량이 달라진다. 즉, 정보의 발생 단계에서는 원천정보가 되고 이것이 변화, 코딩된 정보의 형태로 변환된다. 이때 변환되어 코딩된 정보는 처리단계나(저장단계도 포함해서) 전송도중에 착오가 발생되는 것을 방지하기 위한 대응조치가 취해져야 하는데 정보 이론에서는 채널코딩이라고 한다.

예를 들면 영문자의 경우 원천정보는 문자당 4.75비트가 되지만 실제로는 5비트로 코딩, 기호나 숫자도 포함시키고 착오검출이나(페리티 검출방법 등이 대표적이다) 교신절차(이를 프로토콜이라 한다)에 필요한 비트를 추가시켜 8비트로 변환시킨다.

따라서 1개의 영문자는 실제 정보량이 4.75비트이지만(발생 엔트로피는 이보다 적다 : 88년 4월호 참조) 8비트로 포장되어 3.25비트나 추가되는 셈이다. 8비트를 1바이트(byte) 또는 octatat라 한다. 8비트 정보를 컴퓨터에서 처리할 때는 8개의 버스선을 이용하여 병렬 전송이 되지만 통신로에 전송시킬 때는 2선만 사용하므로 직렬 전송방식으로 바꾸어 주어야 한다. 또한 전송로 잡음을 이겨내게 하고 전송방식에 알맞도록 새로운 코딩방식으로 바꾸어 주기도 하는데 이를 신호변조 과정이라 한다.

한편 저장처리 되는 정보나 통신로에 전달된 정보는 목적지에 도달, 원래의 정보로 재생되어 나타나게 된다. 이를 표현(display)정보라 한다.

〈그림 1〉 문자 정보의 정보량 비교

		입력정보량	전송정보량		출력정보량	
		원천정보 (단위:비트/자소)	컴퓨터	데이터통신	표현방식	표현정보
자소	한글	4.644	5	8(병렬)	8(직렬)	풀어쓰기	8
	영문	4.75	5	8(병렬)	8(직렬)	풀어쓰기	8
단어	한글	1문자=2.62자소	11.004	13.10	–	–	2바이트형 모아쓰기	16
		1단어⇒문자	33.012	39.30	–	–	2바이트형 모아쓰기	48
	영어	1단어=6자소	26.4	30	–	–	풀어쓰기	48

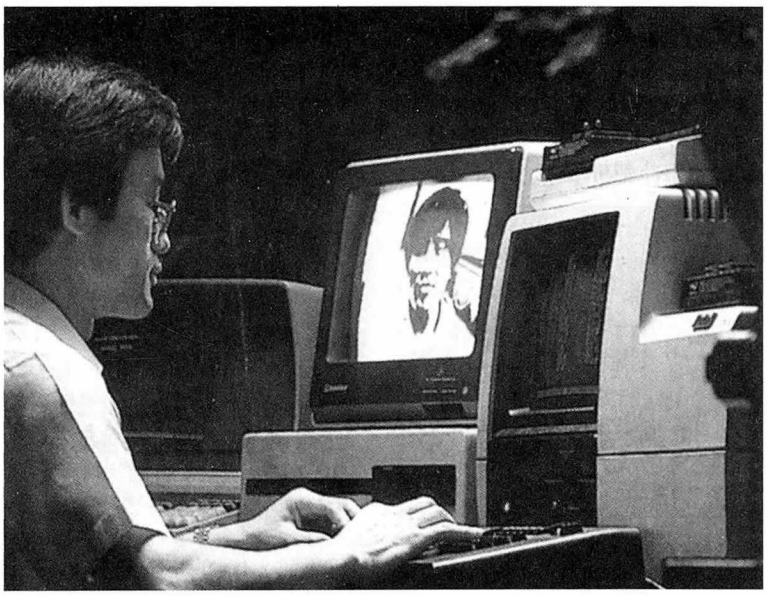
정보를 재생시키는 방법은 문자의 경우, 타자기(프린터기)·CRT모니터 등이고 음성과 스피커 영상인 경우에는 CRT스크린·사진·FAX 등이다. 여기서 기록보존이 가능한 녹음기·타자기·사진·FAX 등을 하드카피, 스크린·모니터 등 기록 보존이 어려운 표현 방식을 소프트카피라고도 한다. 영문 표현정보의 경우 8비트가 되지만 한글의 경우는 양상이 달라진다.

2. 한글의 출력 정보량

한글의 원천정보는 4.644비트이며 전송로에 전송되는 정보량은 영문의 경우처럼 8비트로 계산할 수가 있다. 또한 한글은 2.62자소가 1문자를 이룬다. 따라서 1문자당 정보량은 2.62자소×4.644비트=11.004비트가 되고 5비트로 코딩되므로 13.10비트가 된다. 한글은 모아쓰기라는 독특한 표현방식을 써야 하므로 풀어쓰기 입력에 모아쓰기 출력방식을 채택해야 한다. 따라서 풀어쓰기로 입력된 정보를 모아쓰는 과정을 첨가하고 2000여자에 달하는 모아쓴 문자를 표현하자면 어드레스를 지정하는데 8비트로는 어려움이 따른다. 때문에 2바이트 16비트로 된 조합형 코드를 사용해야 한다. 이는 영문자에 비해서 다소 불편한 점이 있다. 그러나 이를 조합형 2바이트 코드로 나타낸 다고 볼 때는 16비트가 되어 오히려 유리한 점이 되기도 한다. 이는 풀어쓰기와 모아쓰기의 과정에서의 정보압축의 효과로 생각된다. 한글의 경우 1단어를 3자소로 구성되어 있다고 보면(아직 단어당 자소 구성에 대한 정확한 조사없음), 한 단어를 기준해 볼 때 원천정보로서는 33.012비트(39.30비트)이다. 따라서 2바이트 완성형 코드로 출력시킬 때는 영어의 경우와 같이 1단어당 3자소×16비트=48비트가 됨

을 알 수 있다.

이것은 한글은 영어처럼 입력과정과 전송과정에서 동일한 과정을 거치지만 출력표현 과정에서 모아쓰기라는 과정을 거치기 때문에 다소 복잡해지는 것을 의미한다. 그러나 1단어가 영어의 경우 6자소에 비해 1.86자소가 더 많은데도 동일한 정보량으로 표현되는데 이는 출력표현상에서 모아쓰기 알고리즘이 추가된다거나 문자발생장치의 용량 증가가 수반되는 단점을 보완하고도 남을 만큼 정보능력상 우월한 위치에 있기 때문이다. 모아쓰기가 단순히 영어와의 직접 호환성의 관점에서 보면 불편하고 속도가 느린 것처럼 보이나 실제 정보의 표현에서는 우월한 위치에 있음을 알 수 있다.



3. 음성 정보량의 계산

음성 정보의 경우 정보량을 계산해 보자. 음성정보는 대개 4KHz 이내에 그 에너지가 집중되어 있다. 그리하여 1Hz당 2개의 표본을 취하고(이를 나이키스트의 표본화비라 한다) 한 표본당 256등분 준위(quantizing ; 量子化)로 가른다면 한 표본은 8비트의 정보량을 지닌 것이 된다. 따라서 음성대역 4KHz 전역에 걸쳐서는 8000개의 표본이 있어야 하므로 총 정보량은 8000개×8비트=64000비트/sec가 된다.

이와 같은 방식을 PCM(Pulse Code Modulation : 펄스코드변조)이라 하고 이는 연속적인 애널로그 음성정보가 디지틀 음성정보로 변환 되었다고 할 수 있다.

그렇다면 애널로그 음성정보는 얼마만한 원천정보를 가지고 있을까? 음성으로 주고받는 말은 1초에 12음소가 유통된다. 따라서 문자기준으로 보면 12자소×5비트=60비트 정도의 정보량이 이동하는 셈이된다. 결국 음성으로 이야기 한다는 것은 초당 50~100비트 정도의 정보량을 가지는 셈이다. 이로 볼 때 음성정보가 비효율적이며, 문자정보가 얼마나 압축효과가 큰지 알만하다. 따라서 64Kbps의 음성정보는 실제의 정보를 표현하는데는 모두 기여하지 못하고 군더더기에 불과하다는 것이다(각국 언어의 경우 모두 비슷하다). 이와 같은 군더더기를 없애주는 방법이 음성 압축기술이며, 적응차동, PCM 등이 있다. 음성정보의 종국적 최대 압축 방법은 음성을 인식한 후 문자코드로 바꾸어 전송했다가 다시 음성으로 합성시키는 방식일 것이다.

음성은 이와 같이 군더더기가 많은 정보이지만 필기를 한다든가 그림을 그릴 필요없이 인간대 인간의 직접 대화가 가능하다는 점이 큰 특징이며, 손쉽게 저장되지 못한다는 단점을 아울러 가지고 있다. 이로 볼 때 인간의 청각 기능은 64Kbps 정도의 정보를 50~100bps정도로 압축시켜 인지할 수 있는 장치라는 것을 알 수 있다.

4. 영상 정보량의 계산

도형과 영상정보를 인식하는 우리의 시각 구조는 대개 1억개 센서로 구성되었다고 한다. 1억개의 포인트에서 만약 명암 정도를 16단계로 구분한다면 한점에서는 4비트의 정보가 필요하므로 전체적으로 4억비트가 되고, 다시 천연색으로 볼 때 다시 2비트가 더 필요하다고 가정하면 8억비트가 된다. 또한 움직이는 영상까지를 포착하는 능력을 1초당 30화면이라고 본다면 실로 240억 bps의 정보를 처리하는 정보기기임을 알 수 있다. (이를 다시 수정체의 렌즈에서 2차원 FFT 한다고 볼 때 놀라운 정보압축 능력을 가지고 있다.)

실로 우리가 다루는 시각 정보기구인 TV에 대한 정보 취급량을 계산해 보자. TV에서는 525개의 점으로 구성된 525개의 주사선으로 1화면이 구성되며 초당 30장면을 처리한다. 한점에서 명암정도 차이가 16등분 된다면 4비트의 정보를 가지게 된다. 따라서 525×525×4bit×30화면/초=

31,875,000bps의 정보량이 처리되는 셈이다. TV 방송을 행할 때는 잡음에 비해 신호의 크기가 1000배(30db) 이상이 되어야하므로 소요대역폭은 3.187MHz가 된다.

> 통신로 용량을 나타내는 수식은 $C=Bw \log_2(1+s/n)$으로 표현한다(88년 5월호 참조). 따라서 TV는 31,875,000bps의 통신로 용량을 가진 경우가 된다. 또한 s/n이 30dB이면 $\log_2(1+1000)$이 된다. 계산의 편의상 대수항을 $\log_2 1024=10$으로 가정하면 영상 대역폭은 Bw는 3.187 MHz가 된다.

이와 같은 이유로 TV 1채널은 영상대역폭을 4MHz로 잡는다. 앞에서 음성대역을 4KHz로 잡았으므로 음성에 비해 1천배의 대역폭이 필요함을 알 수 있다.

"百聞不如一見"이라는 말은 정보량에서는 "千聞不如一見"으로 수정되어야 할 것이다.

물론 천연색인 경우는 이보다 정보량이 더 많아지고, 정보량이 많아질수록 기기는 복잡해지고 몸체가 커지게 된다. 전화기와 TV수상기를 비교해 보면 쉽게 납득할 수 있다. 또한 영상정보기기는 평면적인 2차원인데 비해서 음성정보는 1차원 정보형태를 가진다. 영상을 처리할 때 스케이닝이라는 2차원 표현방식이 필요한 것도 이러한 특징에서 유래된다. 컴퓨터에서 처리하는 정보가 초기의 데이터 처리에서 음성처리 그리고 영상처리로 넘어가는 것도 이와 같은 정보량의 차이에서 기인되었음을 알 수 있다.

5. 정보의 압축기술과 정보량의 변화

앞에서 서술한 것처럼 영상정보는 데이터 정보나 음성정보에 비해서 월등한 정보량을 가지고 있으므로 이를 처리하거나 전송하는데 있어 막대한 비용이 소모된다. 이 때문에 음성정보처럼 영상정보에서 압축기술이 연구되고 있다. 변환방법(Transform Method), 백터양자화(Vector Quantizing)기법 등이 그것이다. 이와 같은 압축기술은 음성정보와 마찬가지로 군더더기율을 제거하는 것이 근본 원리이다. 그러나 최근의 압축기술 동향은 이러한 군더더기 제거 이외에도 인간의 시각과 청각정보 처리 기능의 특성을 이용하는 쪽으로 진행되고 있다. 왜냐하면 군더더기를 제거하는 과정에서 자연성(Naturality)도 사라지기 때문이다. 한 예를 들자면 앞에서 잠깐 언급했던 바와 같이 음성을 인식하여 문자코드화 하고 개개인의 음성특징(이를 聲紋識別 기술이라 한다)을 파라메터로 코딩하여 전송했다가 다시 음성으로 재생하고, 특징을 파라메터를 추가한다면 음성정보의 압축도 달성하고 자연성도 재생할 수 있을 것이다. 그러나 이와 같은 과정은 아직 달성하기 어려운 단계에 있으며, 영상정보는 더욱 더 어려운 과제로 좀더 시간이 지나야 하며 보다 고차원의 기술진보가 요구된다. 상세한 압축기술에 대해서는 차후에 다시 언급하겠지만 여기서 알아두어야 할 점은 정보를 압축한다 해도 원천정보량 이하는 불가능하며, 자연성을 해치지 않으려면 또다시 추가적인 정보량이 필요하다는 것이다.

6. 결어

지금까지 정보의 단계에 따라 달라지는 정보량에 대해서 살펴 보았다. 원천정보(입력) 처리·전송정보 및 출력 정보량에서 단계를 지날 때마다 자꾸 포장되고 있음을 살폈다. 한글의 원천정보가 4.644비트에서 5비트로 되는 것은 4.644비트로 처리하는 것이 불가능하고, 다시 8비트로 되는 것은 착오방지를 목적으로 비트가 추가되기 때문이다. 출력정보에서는 원천정보로 다시 복원되는 과정에서 추가된 정보는 사라진다. 이와 같은 단계변화에서 정보량이 달라진다면 기준량을 어디에 맞추어야 할 것인가. 영문의 경우 전과정에서 1자소에 8비트이므로 1단어 6자소의 구성으로 볼 때 48비트로 계산하는 방식이 표준방식이다. 그러나 한글의 경우에는 1단어를 7.86자소로 본다면 전송정보는 7.86자소×8비트=62.88비트가 되나 2바이트 완성형 코드를 사용할 때는 출력에서 48비트로 줄어들어 영어와 동일하게 된다. 결국 풍부한 어휘 능력과(영어에 비해 14.88비트가 앞선다) 고도의 압축능력이 있음을 알 수 있다. 따라서 영어와 같이 8비트 기준과 단어당 48기준으로 삼아도 무방할 것으로 보인다. 물론 1단어당 몇 자소로 구성되어 있는지의 정밀조사가 진행되어야 하고 이때는 좀더 감소될 가능성이 많다. 또한 1문자가 몇개의 자소로 구성되든 2바이트 코드인 경우 16비트인 것은 변함 없다. 이와 같은 한글의 정보능력은 동양적인 사고에서 발생하였으면서도 서양적인 정보기구에도 훌륭하게 적용될 수 있도록 창제되어 있음을 알 수가 있다. 얼마나 다행스러운 일인가? 그러면서도 한글의 정보처리 속도가 떨어진다든가 불편하다는 편견은 전체를 살피지 못하고 부분적인 관찰에 지나지 않는 단견임을 깨달아야 한다. ♣