한국음향학회지 제22권 제6호 pp. 443~451 (2003)

1차원 SPIHT를 이용한 가변 비트율 음성 부호기의 설계

Design of a Variable Bit Rate Speech Coder Based on One-dimensional SPIHT

나 훈*, 정 대 권*
(Hoon Na*, Dae-Gwon Jeong*)

*한국항공대학교 항공전자공학과

(접수일자: 2003년 3월 19일; 수정일자: 2003년 7월 7일; 채택일자: 2003년 7월 28일)

코드북 가반의 CELP 부호기는 코드북에 미리 할당된 부호화 비트율에 따라서 여기 신호를 모델링한 후 코드북을 이용하여 음성신호를 합성한다. 따라서 임의와 다양한 비트율을 하나의 부호기에서 지원하지 못하는 단점이 있다. 본 논문에서 제안하는 가변 비트율 부호기는 웨이블렛 변환 (wavelet transform)과 1차원 SPIHT (one dimensional SPIHT)를 이용하여 현재 프레임에 할당되는 비트수에 따라서 여기신호를 부호화한다. 또한 CELP 부호기의 경우처럼 특정한 몇 가지 형태로 여기신호 (또는 코드북)를 모델링할 필요가 없고, 정확한 피치정보가 없어도 여기신호를 사용자의 요구에 따라 다양한 비트율로 부호화할 수 있다. 그 결과 코드북이 존재하지 않기 때문에 부호기의 복잡도가 낮으며, CELP 기반의 G.729와 G.723.1 부호기와의 음질 비교 결과 동등하거나 나은 결과를 보여준다.

핵심용어: 음성 부호기, CELP, 웨이블렛, 가변비트율
투고분야: 음성처리 분야 (2.2, 2.4)

Since a codebook-based CELP coder models its excitation signal according to one of several bit rates pre-assigned to codebooks and synthesizes speech signal using codebooks, it can not support encoding of speech signal at an arbitrary bit rate in one encoder. The proposed variable bit rate speech coder encodes the excitation signal based on the bit rate assigned to a present frame of speech using one-dimensional SPIHT and wavelet transform. Also it does't need to model excitation signal (or codebook) to some types as CELP coder, and can encode excitation signal at various bit rates without exact pitch information according to user requirement. As a result, since the coder doesn't have a codebook structure, it has relatively low coder complexity and provides equal or better speech quality compared to G.729 and G.723.1 coder.

Keywords: *Speech coder, CELP, Wavelet, Variable bit rate*
ASK subject classification: *Speech signal processing (2.2, 2.4)*

I. 서 론

정보 통신 사회로의 발전이 가속화되면서 유/무선 채널, 인터넷, 인공 위성등을 이용한 다양한 디지털 통신 방식이 개발되고 실용화되고 있다. 8 kHz로 샘플링된 음성신호는 유선통화품질 (toll quality)로 전송하기 위해서는 비교적 높은 대역폭이 요구된다. 그러므로 이동 통신이나 인터넷 등과 같은 대역폭이 제한되어 있는 환경에서는 대역폭을 큰 폭으로 줄일 수 있는 효율적인 부호기를 설계하는 것이 필수적이다.

현재 널리 사용되고 있는 CELP 부호기는 음성신호를 프레임이라 부르는 일정한 시간 단위로 분할한 후, 각 프레임에 해당되는 필터의 계수와 입력 여기신호를 효율적으로 설계 및 부호화함으로써 원래의 신호를 재생하는 방법이다. 부호기는 성도의 특성을 반영하는 계수를 구하는 LPC (Linear Predictive Coding) 분석과 LPC 분석 후에 프레임을 다시 여러 개의 서브 프레임으로 분할 한 후 각각의 서브 프레임에 대해 유성음의 특성을 고려해서 임펄스의 반복 주기에 해당되는 피치 정보를 찾는 LTP (Long Term Prediction) 분석, 그리고 LTP 필터의 입력에 해당되는 여기신호를 모델링하는 부분으로 구성되어 있다. 여기신호를 모델링하는 방법으로 코드북 (codebook) 방식을 사용하며, 코드북은 유성음 성분을 모델링하기 위한 적응 코드북 (adaptive codebook)과 무성음 성분을 모델링하는 고정 코드북 (fixed codebook)으로 나누어진

책임저자: 나훈(nahoon@mail.hangkong.ac.kr)
412-791 경기도 고양시 덕양구 화전동 200-1
한국항공대학교 항공전자과 대학원
(전화: 02-300-0055)

다[1-4].

CELP에 사용되는 코드북은 코드북에 할당된 비트율에 따라서 여기신호를 모델링하여 설계하기 때문에 고정 비트율만을 지원하게 되므로 다양한 비트율을 한 부호기에서 동시에 지원하지 못 하는 단점이 있다. 또한 낮은 비트율에서는 부호화 과정에 필요한 비트수가 전체적으로 줄어들게 되므로 합성된 음성신호의 음질이 부호화 각 과정의 오차에 따라 민감하게 변화한다.

본 논문에서 제안하는 가변 비트율을 갖는 음성 부호기는 여기신호를 부호화 할 때 웨이블렛 변환 (wavelet transform)과 1차원 SPIHT (one dimensional SPIHT)을 이용하여 현재 프레임에 할당되는 비트수에 따라서 부호화하기 때문에 CELP의 경우처럼 특정한 몇 가지 형태로 여기신호 (또는 코드북)를 모델링할 필요가 없으며, 코드북이 존재하지 않기 때문에 코드북 탐색 과정이 생략되어 부호기의 복잡도가 높지 않다. 또한 여기신호 부호화 과정에서 사용자의 요구에 따라 다양한 비트율로 부호화가 가능한 구조를 가지고 있다.

본 논문은 모두 5장으로 구성되며 내용은 다음과 같다. 먼저 제1장에서는 연구를 하게 된 배경 및 필요성과 연구 내용에 대해 기술하고, 제2장에서는 제안한 부호기의 기본 구조와 비슷한 CELP 부호화기에 대한 내용을 설명하였다. 제3장에서는 제안하고 있는 가변 비트율을 갖는 음성 부호기에 관한 내용을 설명하고 제4장에서 기존의 CELP 부호기와 제안한 부호기에 대한 성능 비교를 하였으며, 마지막으로 제5장에서 결론을 맺었다.

II. CELP 음성 부호화기

CELP 음성부호기의 음성 분석 알고리즘은 아래의 그림 1과 같이 크게 3개의 블록으로 나누어진다.

A 블록에서는 먼저 음성신호에 창 함수 (window function) $W(z)$를 취하여 음성신호를 20-30 ms의 프레

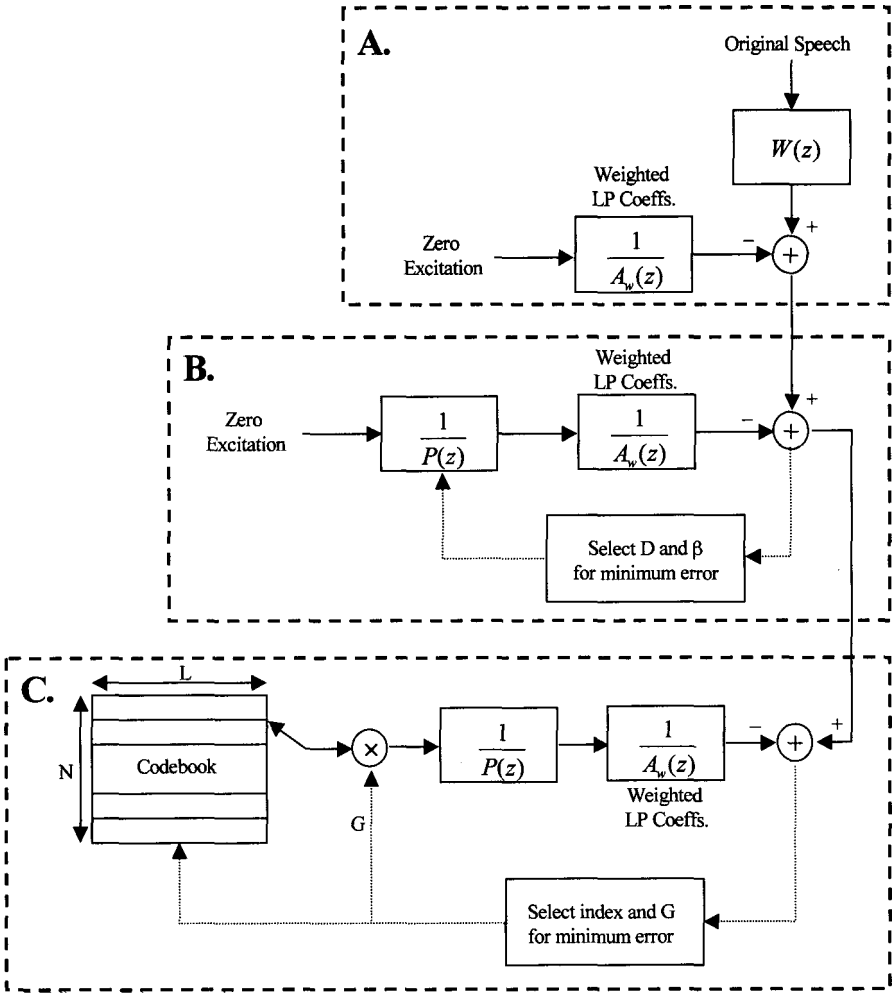


그림 1. CELP 부호기의 블록 다이어그램
Fig. 1. Block diagram of a CELP coder.

임 단위로 나눈다. 그 후 음성신호의 스펙트럼 포락선을 모델링하는 단구간 예측기 (STP: Short-Term Predictor)의 $A_w(z)$의 필터 계수를 구하기 위해서 LPC 분석을 수행한다. 그 결과로 얻어진 필터계수를 적용한 복원신호과 원신호와의 차신호를 B블록에 전달한다. B 블록은 A 블록에서 얻어진 LPC 계수를 이용하여 복원된 음성신호의 유성음 성분에 관련된 피치 정보를 구하기 위한 장구간 예측 (LTP: Long-Term Prediction)을 수행한다. 즉 LTP 분석을 통해 반복적으로 잔차신호를 줄여나가는 방법으로 필터의 계수들을 최적화한다. 구체적으로 구해진 잔차신호를 몇 개의 부 프레임 (sub-frame)으로 나눈 후에 각각에 부 프레임에 대해서 개루프 방법 (open-loop method)과 폐루프 방법 (closed-loop method)을 순차적으로 적용하여 지연값 (delay)과 스케일 인자 (scale factor)를 구한다. C 블록은 LTP 필터의 입력에 해당되는 여기신호를 결정한다. 일반적으로 여기신호는 랜덤 백색 가우스 시퀀스 (random white Gaussian sequence)로 구성된 코드북에서 선택한다. 최적의 여기신호를 찾기 위해서 코드북의 각 벡터들로부터 합성된 음성을 생성하고, 그 후 합성된 음성 (synthesized speech)과 본래의 음성 (original speech) 신호와의 오차를 계산한다. 이 오차를 최소로 하는 벡터가 여기신호로 선택된다[1-4].

III. 가변 비트율을 갖는 음성 부호화기

3.1. 개요

CELP 기반 부호기는 낮은 비트율의 부호기일수록 부호기의 부호화 과정 각 단계마다 파라미터를 부호화 할 수 있는 비트수가 줄어들기 때문에 높은 비트율의 부호기보다 합성된 음성신호의 음질이 파라미터로 인해 발생한 오차에 따라 민감하게 변화한다. 따라서 기존의 CELP 부호기의 단점을 고려할 때 다음과 같은 조건을 만족하는 부호기가 필요하다.

(1) 부호기는 피치 정보에 영향을 적게 받으면서 유성음 구간과 천이 구간을 효과적으로 모델링할 수 있어야 한다.

(2) 다양한 잔차 (residual)신호의 형태를 지원하기 위해 고정된 코드북 형태로 잔차신호를 모델링하지 말아야 한다.

(3) 가변 비트율을 지원하도록 하기 위하여 동일한 부호화 방법을 적용하고, 사용자의 용도와 채널 상태에 따른 비트율 선택을 가능하도록 부호기를 설계해야 한다.

3.2. 부호기 구조

그림 2는 제안한 부호기의 블록 다이어그램을 나타낸다.

아날로그 음성신호는 8 kHz로 샘플링된 후에 16 비트 선형 PCM 부호기를 통과하여 부호기의 입력 음성신호 $s(n)$으로 변환된다. 부호기는 음성신호를 30 msec (240 samples)의 프레임으로 단위로 분할하여 필터링, LPC 분석 등의 부호화를 수행한다. 따라서 다음 프레임은 30 msec 천이하여 얻게 된다. 음성 프레임은 고역 필터 (HPF)를 통과하며 음성신호의 직류 성분이 제거된 후 해밍창 (Hamming window)함수를 통과하여 4개의 서브 프레임으로 분할된다. 분할된 각각의 서브 프레임에 대해서 10차의 LPC 분석을 수행하여 단구간 예측필터의 LPC 계수를 구한다.

LPC 분석을 하여 최적의 LPC 계수를 이용하여 음성신호를 LPC 역 필터링함으로써 여기신호를 구할 수 있다. 여기신호의 웨이블렛 변환 형태를 결정하기 위해, 여기신호의 유/무성음을 결정한 후, 유/무성음의 여부에 따라 웨이블렛 변환을 수행한다. 일반적으로 유성음은 저주파 대역의 에너지가 크고 무성음은 고주파 대역의 에너지가 크므로 웨이블렛 변환시에 이를 고려하여 변환을 수행하면 효과적인 웨이블렛 계수의 부호화가 가능하다. 웨이

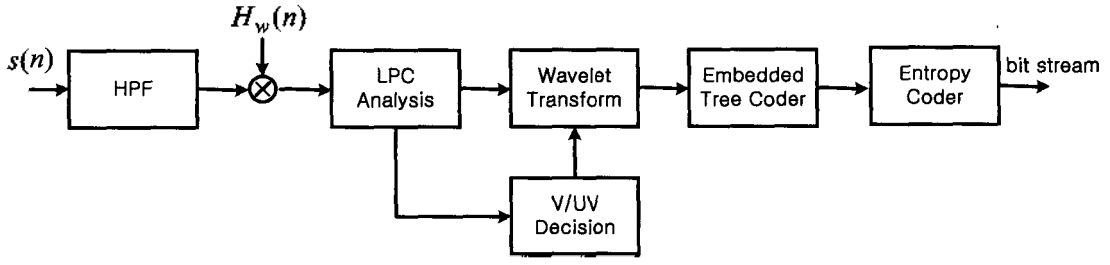


그림 2. 제안한 부호기의 블록 다이어그램
Fig. 2. Block diagram of proposed encoder.

블랫 변환으로 얻어진 계수들을 부호기에서 정한 비트율에 따라 임베디드 트리 구조를 이용하여 부호화한 후 엔트로피 부호화를 거쳐 구해진 비트열을 전송한다.

3.2.1. LPC 분석

LPC 분석은 제 2장에서 설명한 기존의 CELP 기반의 부호기와 동일한 방법을 사용하여 수행된다[2,16,17]. 30 msec의 음성 프레임 $s(n)$은 직류 성분을 제거하기 위해서 다음 식과 같은 전달 함수를 갖는 고역필터 (HPF)를 사용한다.

$$H_{HPF}(z)=\frac{1-z^{-1}}{1-\frac{127}{128}z^{-1}} \tag{1}$$

따라서 직류 성분이 제거된 음성신호는 $s_{HPF}(n)$와 주파수응답 $S_{HPF}(z)$는 다음 식과 같이 표현된다.

$$S_{HPF}(z)=H_{HPF}(z)S(z) \tag{2}$$

$$s_{HPF}(n)=s(n)-s(n-1)+\frac{127}{128}s_{HPF}(n-1) \tag{3}$$

LPC 분석을 수행하기 전에 음성 신호의 천이 구간의 정보를 효과적으로 반영하기 위해 인터폴레이션 구조를 갖도록 LPC 해석 프레임을 구성한다[2]. 본 논문에서는 G.723.1 부호기에 사용된 것과 같은 구조의 해석 프레임을 사용하였다. 다음과 같이 과거의 음성 프레임 120 샘플과 현재의 음성 프레임 240 샘플로 새로운 360 샘플 길이를 갖는 LPC 해석 프레임을 그림 3과 같이 구성한다. LPC 해석 프레임 $s_{LPC}(n)$은 다음 식과 같이 결정된다.

$$s_{LPC}(n)=\begin{cases} s^{-1}_{HPF}(n), & 0\le n\le 119 \\ s^{0}_{HPF}(n), & 120\le n\le 359 \end{cases} \tag{4}$$

여기서 $s^{-1}_{HPF}(n)$와 $s^{0}_{HPF}(n)$는 각각 과거의 음성신호 120 샘플과 현재의 240샘플의 음성신호 $s_{HPF}(n)$를 의미한다.

360샘플의 LPC 해석 프레임은 다음과 같이 표현되는 길이 $L=180$인 해밍 창함수를 이용하여 4개의 LPC 서브 프레임으로 분할된다.

$$H_w(n)=\begin{cases} 0.54-0.46\cos(2\pi\frac{n}{L-1}), & 0\le n\le L-1 \\ 0, & otherwise \end{cases} \tag{5}$$

LPC 서브 프레임은 $s^i_w(n)$은 다음 식과 같이 구한다.

$$s^i_w(n)=s^i_{LPC}(n)H_w(n),\quad 0\le i\le 3,\quad 0\le n\le L-1 \tag{6}$$

여기서, 각각의 서브프레임은 인접하는 서브프레임과 120샘플이 서로 중첩되도록 다음과 같이 만들어진다.

$$s^i_{LPC}(n)=s_{LPC}(n),\quad i\times 60\le n\le (L-1)+60\times i,\quad 0\le i\le 3 \tag{7}$$

각각의 LPC 서브 프레임 $s^i_w(n)$에 대해서 10차의 LPC 분석을 수행하여 LPC 계수를 구한다. LPC 분석에 사용되는 필터의 전달함수는 다음 식과 같다.

$$A_i(z)=\frac{1}{1-\sum_{j=1}^{10}a_{ij}z^{-j}},\qquad 0\le i\le 3 \tag{8}$$

3.2.2. 유/무성음 결정

LPC 분석에서 최적의 LPC 계수를 구한 후에 다음의 일련의 식과 같이 $s^i_w(n)$에 역 필터링을 취하여 여기신호의 주파수응답 $R_i(z)$를 구할 수 있다.

$$S^i_w(z)=A_i(z)R_i(z),\qquad 0\le i\le 3 \tag{9}$$

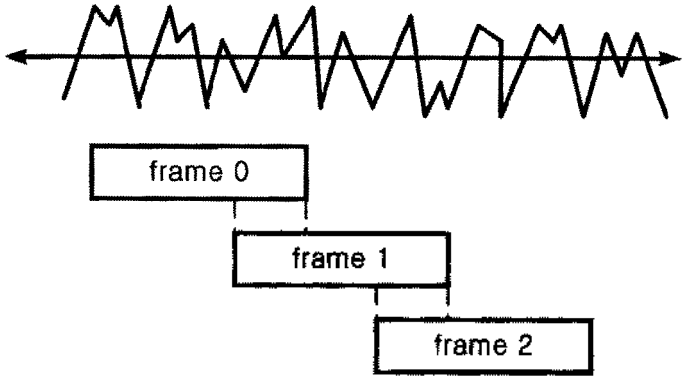


그림 3. LPC 해석 프레임의 구조
Fig. 3. LPC analysis frame scheme.

$$R_i(z) = A_i^{-1}(z) S_w^i(z), \qquad 0 \le i \le 3 \tag{10}$$

여기서,

$$A_i^{-1}(z) = 1 - \sum_{j=1}^{10} a_{ij} z^{-j}, \qquad 0 \le i \le 3 \tag{11}$$

여기신호 $r_i(n)$은 식 (10)을 역 z 변환하여 다음과 같이 구할 수 있다.

$$r_i(n) = s_w^i(n) - \sum_{j=1}^{10} a_{ij} s_w^i(n-j), \quad 0 \le n \le L-1, \quad 0 \le i \le 3 \tag{12}$$

식 (12)에서 구한 여기신호에 대해서 유/무성음 결정을 하기 위해 웨이블렛 변환, 여기신호 에너지, 영 교차율 (zero crossing rate)을 이용한 알고리즘을 사용한다[18].

3.2.3. 여기신호의 웨이블렛 변환

그림 4는 샘플 8개가 한 프레임인 경우 레벨이 3인 웨이블렛 변환을 한 후 그 때의 주파수 분포를 나타낸 것으로 H는 고주파 대역 L은 저주파 대역을 의미한다. 즉 웨이블렛 변환 과정은 신호의 주파수 대역을 계속해서 서로 다른 해상도를 갖는 고주파 대역으로 분해해 가는 과정으로 볼 수 있다

음성신호의 경우는 신호의 특성상 유성음의 경우는 저주파수 대역의 에너지가 크고, 무성음의 경우는 고주파수 대역의 에너지가 상대적으로 크다. 따라서 그림 4에서처럼 일반적인 웨이블렛 변환을 여기신호에 대해 일률적으로 적용하면 유성음의 부호화에는 문제가 없으나 고주파수 대역의 에너지가 큰 무성음의 경우는 계수들의 에너지 분포가 효과적으로 되지 못하기 때문에 양자화시에 더 많은 비트수를 필요로 하게 된다.

본 논문에서는 양자화의 효과적인 적용을 위해서 음성의 여기신호에 대해서 유/무성음 판별을 한 후, 유/무성음에 대해 각각 다른 주파수 분포를 갖는 웨이블렛 변환을 수행하여 양자화의 효율을 높였다. 즉 유성음의 경우는 일반적인 웨이블렛 변환을 수행하고 무성음의 경우는 웨이블렛 패킷 변환 (wavelet packet transform)을 수행하여 주파수 분포를 설정하였다. 그림 5는 웨이블렛 변환 레벨이 3인 경우 유성음과 무성음에 대한 웨이블렛 변환의 주파수 대역 설정 방법을 나타낸다.

3.2.4. 임베디드 트리 부호화

임베디드 트리 부호화는 웨이블렛 변환에 따른 계수들의 공간적, 주파수 유사성을 이용하여 부호화하는 것으로 다음과 같은 특성을 이용한 것이다[11,14,15].

(1) 계수들은 서브밴드 (subband)를 가로질러서 공간적인 자기 유사성을 갖고 있다
(2) 각 서브밴드에서는 각기 다른 해상도 (resolution)를 갖지만 동일한 공간적 위치를 갖는다
(3) 특정 레벨에서의 웨이블렛 계수가 작은 값이면 동일한 공간적 위치에 해당하는 더 낮은 레벨에서는 계수들은 더 작은 값을 갖는다.

(3)의 경우는 유성음의 경우에는 적합하지만 무성음의 경우는 대부분의 에너지가 고주파수 대역에 있기 때문에 적절하지 않으며 3.2.3절에서 제안한 방법을 이용해야 한다. 임베디드 트리 부호화의 대표적인 예로는 Shapiro의 EZW (Embedded Zero-tree coding)[14]와 Said의 SPIHT (Set Partitioning In Hierarchical Trees)[15]가 있다. 두 방식 모두 영상 신호를 부호화하는데 널리 쓰이고 있으며, 본 논문에서는 영상 신호에 적용되는 SPIHT 방식을

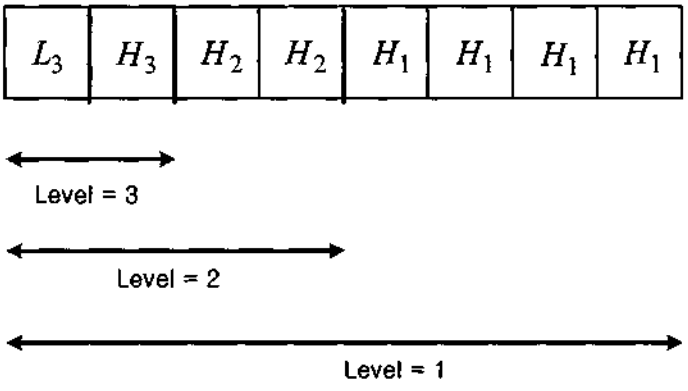
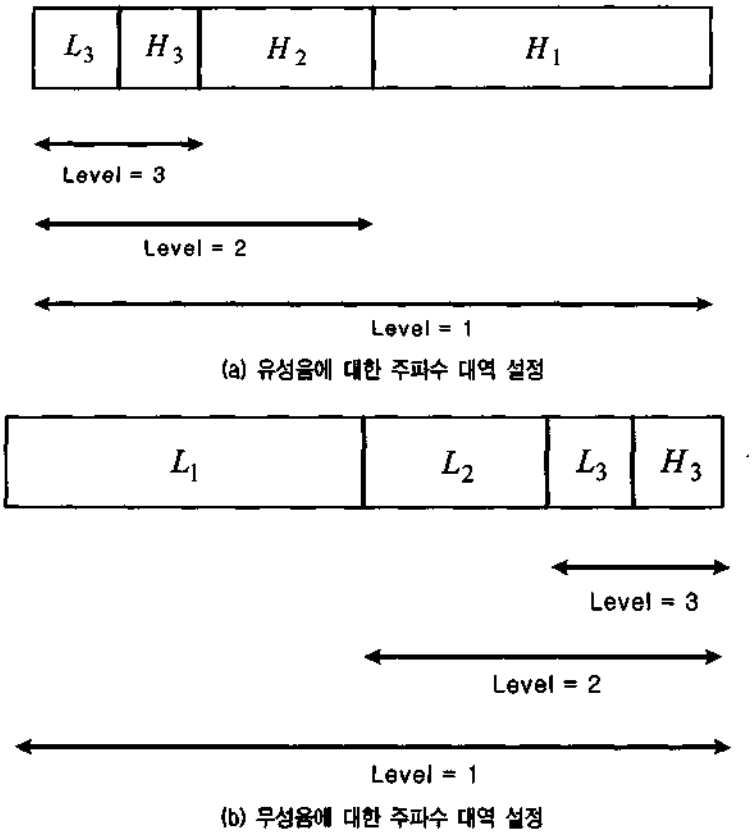


그림 4. 웨이블렛 변환 영역의 주파수 분포
Fig. 4. Frequency distribution of wavelet transform domain.



(a) 유성음에 대한 주파수 대역 설정

(b) 무성음에 대한 주파수 대역 설정

그림 5. 유/무성음에 대한 주파수 대역 설정
Fig. 5. Frequency bands decomposition of V/UV.

음성의 여기신호 부호화에 적합하도록 개선한 MoSPIHT (Modified one dimensional SPIHT)을 제안하였다.

3.2.4.1. MoSPIHT (Modified one dimensional SPIHT)

음성의 여기신호는 EZW나 SPIHT이 적용된 영상 신호와 달리 1차원 공간을 갖는다. 그림 6의 공간 좌표 트리 구조는 웨이블렛 변환 레벨이 2인 경우에 해당된다. 그림에서 ☆는 자손이 없는 노드를 의미하고 A는 자손이 있는 부모 노드, A', A" 등은 자손을 의미한다.

그림 6의 경우처럼 SPIHT의 공간 좌표 트리의 구성은 동일한 해상도를 갖는 같은 레벨의 서브밴드나 다른 레벨의 서브밴드의 구별없이 저주파수 대역의 계수들이 고주파수 대역 계수의 부모가 되며 다음 식과 같은 관계를 갖는다. (i)는 계수의 좌표를 의미하여 $O(i)$는 (i) 바로 아래 자손들의 집합에 해당된다.

$$O(i)=\{(2i),(2i+1)\} \tag{13}$$

그러나 동일 서브밴드 내에서 이러한 트리 구성은 동일 서브밴드 내에서의 공간적인 해상도의 동등성을 위반하며 또한 여기신호의 프레임 사이즈가 클수록 최상위 레벨에서 자손이 없는 뿌리 노드의 수가 증가하게 되어 부호화시에 효율성을 떨어뜨린다. 본 논문에서는 동일 서브밴드 내에서 공간적 해상도의 동등성을 일치시키고, 자손이 존재하지 않는 뿌리 노드가 없는 트리 구조를 사용하였다. 그림 7은 MoSPIHT에서의 공간 좌표 트리 구조를 나타낸다.

MoSPIHT의 경우 자손들의 집합 $O(i,j)$는 그림 7로부터 다음 식과 같은 관계를 갖는다

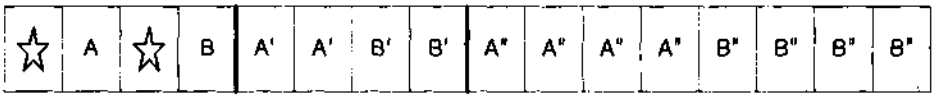


그림 6. SPIHT에서의 여기신호의 공간 좌표 트리 구조
Fig. 6. Spatial orientation tree structure of residual signal in SPIHT.

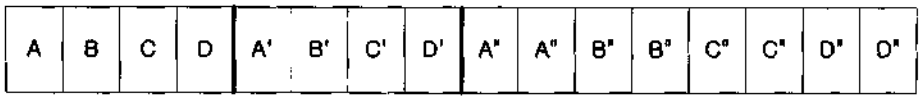


그림 7. MoSPIHT에 대한 유성음 여기신호의 공간 좌표 트리 구조
Fig. 7. Spatial orientation tree structure of voiced residual signal in MoSPIHT.

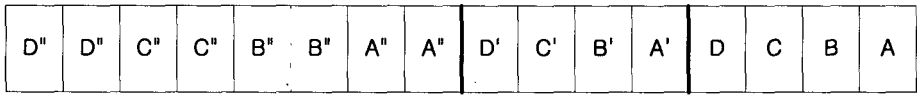


그림 8. MoSPIHT에 대한 무성음 여기신호의 공간 좌표 트리 구조
Fig. 8. Spatial orientation tree structure of unvoiced residual signal for MoSPIHT.

$$O(i)=\begin{cases}\{(2i)\}, & \textit{the highest level}\\ \{(2i),(2i+1)\}, & \textit{otherwise}\end{cases} \tag{14}$$

이상의 방법은 여기신호의 에너지가 저주파수 대역에 분포하는 유성음의 경우이며 에너지가 고주파수 대역에 분포하는 무성음인 경우는 효과적인 부호화를 위해서 3.2.3절에서 제안한 방법으로 웨이블렛 변환을 수행하며 MoSPIHT의 공간 좌표 구성은 다음 그림과 같다.

MoSPIHT과 같은 트리 부호화는 비트 평면상에서 각 비트들을 MSB (Most Significant Bit)에서 LSB (Least Significant Bit)순서로 전송하며 부호화하는 것으로 가장 수치가 큰 계수에 의해 전체적인 비트수의 레벨이 결정된다. 따라서 여기신호를 트리 부호화하기 전에 가장 최상위 웨이블렛 계수들의 평균을 구한 후 이를 뺀 값을 부호화하면 좀 더 효과적으로 부호화에 필요한 비트를 감소할 수 있으며 웨이블렛 계수 C_i는 다음 식과 같다.

$$C_i=\begin{cases}C_i-m, & (i)\in \textit{the highest level}\\ C_i & \textit{otherwise}\end{cases} \tag{15}$$

여기서 $m=\frac{1}{N}\sum_{i=0}^{N-1}C_i$이며 N은 최상위 저주파 대역 레벨에 존재하는 계수의 개수를 의미한다.

3.2.5. 비트 프레임 형식

제안한 부호기는 가변 비트율을 지원하므로 부호기의 초당 비트율을 K bps로 설정했을 때 프레임당 비트 수, N은 다음 식과 같으며 구체적인 비트 할당은 표 1과 같다.

$$N=K_{bps}\times 30_{msec} \tag{16}$$

mean은 식 (15)에서 구한 최상위 저주파 레벨의 평균인 m을 의미하며, $Q_0\sim Q_3$는 임베디드 트리 부호화에 필요한 비트수를 의미한다. 예를 들면 8 kpbs로 부호화하는 경우 $N=240$, $Q_0\sim Q_3=43$이 된다.

3.3. 복호기 구조

그림 9는 복호기의 블록 다이어그램을 나타낸다. 부호기로부터 전송된 비트 스트림은 엔트로피 복호기를 거친 후, 유/무성음 판별 정보에 따라 임베디드 트리 복호기와 역 웨이블렛 변환을 유/무성음에 따라 각각 다르게 적용되어 LPC 합성기를 거치면 합성된 음성신호가 생성된다.

IV. 모의 실험 결과

본 논문에서 제안한 부호기의 성능을 CELP 기반의 ITU-T G.723.1[16]과 G.729[17] 부호기를 사용하여 동일한 비트율에서 비교하였고 시험에 사용된 음성 샘플은 표 2와 같다. 음질의 평가는 SegSNR (Segmental Signal-to-Noise Ratio)를 이용하여 측정하였다.

SegSNR은 30 msec에 해당하는 음성 240 샘플을 한

표 1. 비트수 N에서 제안한 부호기의 비트 할당
Table 1. Bit allocation of proposed encoder at N bits.

Parameter	subframe 0	subframe 1	subframe 2	subframe 3	Total
LP Coeff.					24
V/UV info.	1	1	1	1	4
mean	10	10	10	10	40
Q	(N-68)/4	(N-68)/4	(N-68)/4	(N-68)/4	N-68
Total					N bits

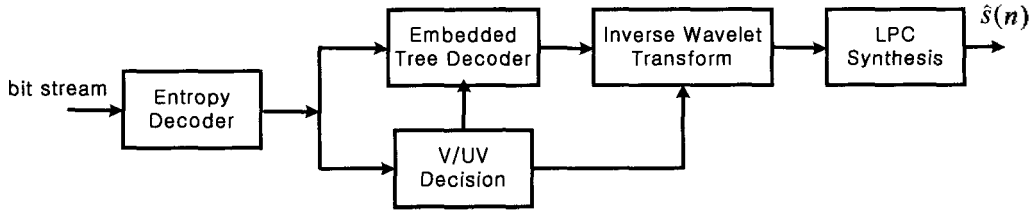


그림 9. 제안한 복호기
Fig. 9. Block diagram of proposed decoder.

프레임으로 하여 다음의 식을 이용해 측정하였다.

$$SNR = 10\log_{10}\left\{\frac{\frac{1}{240}\sum_{0}^{239}s^2(n)}{\frac{1}{240}\sum_{0}^{239}(s(n)-\hat{s}(n))^2}\right\}\ dB \quad (17)$$

$$SegSNR = \frac{1}{M}\sum_{n=0}^{M-1}SNR_i\ \ dB \quad (18)$$

여기서 $s(n)$은 원 음성신호이며 $\hat{s}(n)$은 음성 부호기를 이용하여 부호기에서 합성된 신호를 의미한다. SNR_i는 i번째 프레임에서 신호대 잡음비를 의미하며 M은 묵음 구간을 제외한 프레임 수를 의미한다. 묵음 구간 (silence region)은 SNR_i 계산시 신호 값이 측정되어 SegSNR에 영향을 주므로 묵음 구간으로 판정되면 0 dB로 설정하였다.

4.1. G.729 부호기와 비교 결과

표 3은 8 kbps에서 G.729와 제안한 부호기와의 SegSNR 비교 결과를 나타낸다. 제안한 부호기가 G.729에 비해 SegSNR이 더 나은 결과를 나타내었다.

표 2. 시뮬레이션에 사용된 음성신호 문장
Table 2. Test sentences for simulation.

	sentence	no. of frame
male 1	이번 겨울은 예년과 달리 포근합니다	86
male 2	개인 통신 시대가 조만간에 개막될 것입니다	105
female 1	피는 피부 한 껍질 차이입니다	82
female 2	지나친 흡연은 건강을 해칩니다	106

표 3. G.729와 제안한 부호기의 SegSNR 비교
Table 3. Comparison result with G.729 and proposed coder at 8 kbps.

	G.729 (dB)	proposed coder (dB)
male 1	14.002	14.366
male 2	11.435	12.706
female 1	14.186	14.715
female 2	13.873	14.352

4.2. G.723.1 부호기와 비교 결과

표 4는 6.3 kbps에서 G.723.1과 제안한 부호기와의 SegSNR 비교 결과를 나타낸다. SegSNR 결과는 제안한 부호기가 남성의 경우는 우수하나 여성의 경우 약간 감소하는 결과를 나타내며 청취 비교 결과의 경우는 대등한 음질을 나타내었다.

4.3. G.729, G.723.1 부호기와 부호화 속도 비교 결과

표 5는 Pentium III 900 MHz에서 음성신호 문장을 부호화했을 때의 제안한 부호기의 부호화 시간을 기준으로 한 상대적인 부호화 시간을 나타낸다. 제안한 부호기는 코드북 탐색과정이 필요없으므로 부호화 시간이 단축됨을 알 수 있다.

표 4. G.723.1과 제안한 부호기의 SegSNR 비교
Table 4. Comparison result with G.723.1 and proposed coder at 6.3 kbps.

	G.723.1 (dB)	proposed coder (dB)
male 1	13.636	14.123
male 2	10.761	11.358
female 1	14.311	14.232
female 2	14.058	14.029

표 5. G.729, G.723.1 부호기와 제안한 부호기의 상대적인 부호화 시간 비교
Table 5. Relative encoding time to G.729, G.723.1 and proposed coder.

	proposed coder	G.729	G.723.1
male 1	1	3.2	1.3
male 2	1	3.3	1.4
female 1	1	3.1	1.2
female 2	1	3.0	1.3

V. 결 론

본 논문은 가변 비트율을 하나의 부호기에서 동일한 방법으로 지원하는 임베디드 트리 기반의 음성 부호기를 설계하고 이를 구현하였다. 코드북 기반 CELP 부호기는 여기신호를 코드북을 사용하여 설계하며 코드북에 할당된 부호화 비트에 따라서 모델링한다. 따라서 고정 비트율만을 지원하게 되며 다양한 비트율을 한 부호기에서 동시에 지원하지 못 하게 된다. 또한 낮은 비트율에서는 부호화 과정에 필요한 비트수가 전체적으로 줄어들게 되므로 합성된 음성신호의 음질이 부호화 각 과정의 오차에 따라 민감하게 변화된다. 그 결과 LTP 과정에서 부적절한 피치 정보를 선택하게 되면 피치 정보를 기반으로 동작하는 적응 코드북 또한 여기신호를 잘못 생성하게 되므로 유성음 구간이나 천이 구간을 효과적으로 모델링하지 못하게 되어 음질을 떨어뜨린다.

기존 부호기의 문제점을 개선하기 위해서 본 논문에서 제안한 부호기는 첫째 피치 정보에 영향을 적게 받으면서 유성음 성분의 여기신호를 효과적으로 모델링하고, 둘째 다양한 여기신호 형태를 지원하기 위해서 고정된 코드북 형태로 여기신호를 모델링하지 않으며, 셋째 가변 비트율을 지원할 때 하나의 부호기에서 동일한 알고리즘을 사용한다. 또한 코드북 구조가 아니기 때문에 코드북 탐색 과정을 필요로 하지 않아 복잡도가 높지 않으며 CELP 기반의 G.729와 G.723.1 부호기와의 음질 비교 결과 동등하거나 나은 결과를 나타내었다.

참 고 문 헌

1. L. R. Rabiner, and R. W. Schafer, *Digital Signal Processing of Speech Signal*, Prentice-Hall, 1978.
2. A. M. Kondoz, *Digital Speech*, John Wiley, and Sons, 1996.
3. 안수길, "음성 분석, 모델링," 전자공학회지, 20 (5), May, 1993.
4. 박경범, 음성의 분석 및 합성과 그 응용, 도서출판 그린, 1997.
5. ISO/IEC 144963-3 Subpart 2 *Parametric coding*, 1997.
6. D. W. Griffin, and J. S. Lim, "Mutiband excitation vocoder" *IEEE Trans. on Acoust., Speech and Signal Process.*, **36** (8), 1223-1235, 1988.
7. Digital Voice Systems Inc., *IMBE Vocoder Description*, 1993.
8. K. R. Castleman, *Digital Image Processing* , Prentice- Hall, Inc., 1996.
9. R. K. Young, *Wavelet Theory and Its Applications*, Kluwer Academic Publishers, 1994.
10. C. S. Burrus, R. A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transforms*, Prentice- Hall, Inc., 1998.
11. R. M. Rao, A. S. Bopardikar, *Wavelet Transforms*, Addison Wesley, 1998.
12. S. Kadambe, and G. F. Boudreaux-Bartels, "A comparison of a wavelet functions for pitch detection of speech signals," *Proc. of ICASSP*, 449-452, May 1991.
13. S. Kadambe, and G. F. Boudreaux-Bartels, "Application of the wavelet translation for pitch detection of speech signals," *IEEE Trans. on Information theory.*, **38** (2), 917-924, March 1992.
14. J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. on Signal processing.*, **41** (12), 3445-3461, December 1993.
15. A. Said, W. A. Pearlman, "A new fast efficeint image codec based on set partitioning in hierarchical trees," *IEEE trans. on Circuits and Systems for Video Technology*, **6**, 243 -250, June 1996.
16. ITU-T Recommendation G.723.1, *Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s*, March, 1996.
17. ITU-T Recommendation G.729, *Coding of speech at 8 kbit/s using CS-ACELP*, June, 1995.
18. 나훈, 정대권, "개량형 다중대역 여기 음성부호기의 피치 예측 개선" 한국음향학회지, **20** (3), April 2001.

저자 약력

● 나 훈 (Hoon Na)
한국음향학회지 제20권 제3호 참조

● 정 대 권 (Dae-Gwon Jeong)
한국음향학회지 제20권 제3호 참조