한국음향학회지 제20권 제3호 pp. 34~41 (2001)

개량형 다중대역 여기 (IMBE: Improved Multi-band Excitation) 음성 부호기의 피치 예측 개선

An Efficient Pitch Estimation for IMBE (Improved Multi-band Excitation) Speech Coder

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

*한국항공대학교 항공전자공학과

(접수일자: 2000년 10월 13일; 채택일자: 2001년 2월 27일)

기존의 IMBE (개량형 다중대역 여기: Improved Multi-band Excitation) 음성 부호기의 초기 피치 추정 과정은 전체 부호기 연산 시간의 대부분을 차지하며 또한 미래의 음성 프레임들이 초기 피치 추정시 사용되므로 시간 지연이 유발되어 실시간 구현에 장애 요소로 작용되었다. 또한 무성음에 해당되는 프레임에 대해서도 유성음과 동일한 피치 추정을 수행하므로 알고리즘의 효율성을 떨어뜨린다. 본 논문에서는 초기 피치 추정 전에 다이애딕 웨이브렛 변환 (Dyadic Wavelet Transform)을 이용하여 이를 바탕으로 유/무성음을 판별한 후 유성음으로 결정된 프레임에 대해서만 피치 추정을 행하고 무성음으로 결정된 프레임은 랜덤 잡음을 주어서 부호화시의 처리 시간을 단축하였다. 또한, 초기 피치 추정 전에 판별된 유/무성음을 판별하여 유성음과 무성음에 각기 다른 초기 피치 추정 알고리즘을 사용하고 미래의 두 프레임을 사용하지 않음으로써 송, 수신단에 유발되는 시간 지연을 제거하였다. 그 결과 초기 피치 추정 과정의 상대적인 복잡도가 23% 감소되었고 프레임당 처리 시간이 1/10~1/11로 감소되었고 기존의 부호기와 거의 같은 음질을 얻을 수 있었다.

핵심용어: 음성 부호기, 피치, 웨이브렛
투고분야: 음성처리 분야 (2.4, 2.2)

In an IMBE (Improved Multi-band Excitation) speech coder, initial pitch estimation occupies most of the total computing time for the coder due to complex cost function and exhaustive search over candidate pitches. Future frames in initial pitch estimation cause inevitable time delay. Therefore, it is difficult to implement a real-time coder. Furthermore, unvoiced frames use the unnecessary pitch estimation as in the voiced frames. In this paper, each frame is determined voiced or unvoiced by Dyadic Wavelet Transform (DyWT) and, then, initial pitch estimation is performed only for voiced frame. Therefore different pitch estimation algorithms are employed between voiced and unvoiced frames incurring reduced time delay at transmitter and receiver. Simulation result show that the relative complexity of initial pitch estimation is reduced by 23%, and the processing time decreases down to 1/10~1/11 of the IMBE coder while speech quality is almost maintained.

Keywords: *Speech coder, Pitch, Wavelet*
Ask subject classification : *Speech signal processing (2.4, 2.2)*

책임저자: 나 훈 (nahoon@mail.hangkong.ac.kr)
130-876 서울시 동대문구 휘경1동 172-1호 12통 5반
한국항공대학교 항공전자공학과
(전화: 019-384-7582; 팩스: 02-3159-9986)

I. 서 론

디지털 이동 전화 시스템에 사용되는 음성 부호화기는 4.8kbps (bits per second) 이하의 낮은 전송 속도에서도 현재의 일반 전화기 수준의 음질을 유지할 수 있어야 한다. 64kbps 펄스부호변조방식 (Pulse Code Modulation)이나 32kbps 적응차분 펄스부호 변조방식 (Adaptive Differential PCM)과 같은 파형 부호화 (Waveform Coding) 방식은 부호기의 음질은 좋은 반면 높은 전송 속도와 그에 따른 넓은 대역폭을 필요로 하며, 군사용 목적으로 주로 사용되는 매우 낮은 전송 속도의 선형 추정 부호기 (Linear Predictive Coder)와 같은 매개변수 부호화 (Parameter Coding) 방식은 음질이 좋지 않아 특수 목적 이외에는 사용하기가 곤란하다. 따라서 위의 두 가지 방식의 장점을 모두 갖춘 혼성 부호화 (Hybrid Coding) 방식의 음성 부호기에 대한 연구와 개발이 디지털 이동 전화 시스템의 연구와 개발에 있어서 높은 비중을 차지하고 있다[1-3].

현재 가장 활발히 연구되고 있는 혼성 부호화 방식의 음성 부호기로는 북미 지역의 디지털 이동 전화 시스템인 코드분할 다중접속 (Code Division Multiple Access) 방식에 이용되고 있는 코드여기 선형추정 (CELP: Code Excited Linear Prediction)에 기반을 둔 부호기, 이와 유사한 구조를 갖는 일본 디지털 이동 전화 시스템의 벡터합 여기 선형 추정 (VSELP; Vector Sum Excited Linear Prediction) 부호기, 그리고 개량형 다중대역 여기 (IMBE: Improved Multi-band Excitation)부호화기가 있다.[1,2].

IMBE 음성 부호화기는 한 프레임의 음성에 대한 주파수 영역을 여러 개의 부분 대역으로 분할하여 각 부분 대역에 대해서 유/무성음 판별을 하여 부호화하기 때문에 프레임 단위로 처리하는 CELP나 VSELP 부호화기보다 자연스러운 음성을 합성할 수 있다[1-3]. 그러나, IMBE 부호화기는 적합한 피치를 찾아내기 위해 과거의 두 프레임과 미래의 두 프레임을 이용하므로 기본적인 시간 지연이 생길뿐 아니라 복잡한 피치 추정 과정을 거치기 때문에 부호화시 많은 처리 시간을 요구하게 된다. 또한, 일정한 피치가 존재하지 않는 무성음 구간에서도 유성음 구간과 같은 피치 추정 방법을 사용하여 알고리즘의 복잡도가 더욱 증가하게 된다[4,5].

본 논문에서는 기존의 IMBE 부호기 연산 시간의 90% 이상을 차지하는 피치 추정 과정을 개선하기 위해 유/무성음 판정 알고리즘과 개선된 초기 피치 추정 알고리즘을 제안하고 이를 PC상에서 구현한 후 제안한 방법과 음질을 비교 평가하였다. 이를 위해 첫째, 기존의 IMBE 부호기는 특별한 피치가 필요 없는 무성음의 경우에도 유성음과 동일한 피치 추정을 수행하여 부호기의 복잡도를 증가시키므로, 초기 피치 추정전에 다이애딕 웨이브렛 변환 (Dyadic Wavelet Transform)을 이용하여 유/무성음 판별을 한 후 유성음으로 결정된 프레임에 대해서만 피치 추정을 수행하고, 무성음으로 결정된 프레임은 랜덤 잡음을 주어서 부호화시의 처리 시간을 단축하였다. 둘째, 초기 피치 추정전에 판별된 유/무성음의 여부에 따라 유성음과 무성음에 각기 다른 초기 피치 추정 알고리즘을 사용하고, 끝으로 피치 추정시 미래의 두 프레임을 사용하지 않음으로써 송, 수신단에 유발되는 시간 지연을 제거하였다.

본 논문은 모두 5장으로 구성되며 내용은 다음과 같다. 제2장에서는 IMBE 부호기의 개요를 설명하고 제3장에서는 IMBE 부호기의 실시간 구현을 위해서 제안된 유/무성음 판별 알고리즘과 개선된 피치 추정 방법을 기술하며 제4장에서 기존의 IMBE 부호기와 개선된 부호기와의 실험 결과를 제시하고, 끝으로 제5장에서 결론을 맺는다.

II. IMBE 음성 부호화기

IMBE 음성부호기의 음성 분석 알고리즘은 아래의 그림 1과 같다[3,6].

아날로그 음성신호는 초당 8 kHz로 샘플링되어 이산 데이터로 변환되며 한 프레임의 길이가 20ms, 샘플수가 160인 음성 프레임으로 분할된다. 매 음성 프레임은 고역 필터를 통해서 직류 성분이 제거되며, 저역 필터를 통과한 현재 음성 프레임, 과거의 2프레임과 미래의 2프레임을 이용해서 초기 피치가 추정되며 보다 정확한 피치값을 얻기 위해서 피치 정제가 이루어진다. 또한 피치 정제를 통해서 얻어진 피치 주기에 의해 결정되는 기본 주파수 (fundamental frequency)의 고조파에 대하여 유/무성음

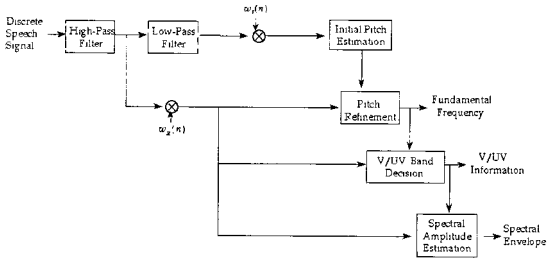

그림 1. IMBE 음성 분석 알고리즘
Fig. 1. IMBE speech analysis algorithm.

판별을 수행하고 판별된 결과에 대한 유/무성음의 주파수 대역에 따른 스펙트럼의 진폭이 결정된다. 이와 같이 각각의 음성 프레임에서 얻어지는 IMBE 부호기의 파라미터는 피치주기, 유/무성음 결정 계수와 스펙트럼의 진폭이다.

2.1. 피치 추정

피치 추정의 목적은 현재의 음성 프레임 $s_w(n)$의 피치 P_0를 추정하는 것으로써 기본 주파수 w_0와 다음과 같은 관계를 가진다.

$$P_0 = \frac{2\pi}{w_0} \tag{1}$$

여기서 P_0는 샘플링된 신호들로부터 추정되며, 일반적으로 인접하는 프레임들의 피치가 급하게 변하지 않도록 현재 프레임의 피치를 결정할 때 과거와 미래의 프레임의 피치를 고려해야 한다.

피치는 초기 피치 추정과 피치 정제라는 두 단계를 거쳐서 추정된다. 초기 피치 추정은 1/2 샘플 정확도로 $\{21, 21.5, \cdots, 121.5, 122\}$ 중 하나의 샘플 거리에 해당되는 값으로 추정되며 초기 피치를 구한 후 피치 정제를 통해서 1/4 샘플 정확도를 갖는 기본 주파수 w_0를 결정하게 된다.

2.2. 초기 피치 추정

초기 피치 추정치를 얻기 위해 식 (2)로 정의되는 오차함수 $E(P)$가 $\{21, 21.5, \cdots, 121.5, 122\}$에 대하여 구하고, 이중 최소의 오차함수 값을 갖는 피치를 초기 피치 $\hat{P}_I$로 결정한다.

$$E(P) = \frac{\sum_{j=-150}^{150} s_{LPF}^2(j) w_I^2(j) - P \cdot \sum_{n=-\lfloor 150/P \rfloor}^{\lfloor 150/P \rfloor} r(n \cdot P)}{\left[\sum_{j=-150}^{150} s_{LPF}^2(j) w_I^2(j)\right] \left[1 - P \cdot \sum_{j=-150}^{150} w_I^4(j)\right]} \tag{2}$$

여기서 $w_I(n)$은 초기 피치 탐색을 위해 사용되는 창(window) 함수이며 $r(t)$는 저역 통과 필터를 거친 음성 신호 $s_{LPF}(n)$의 자기상관 함수 (autocorrelation function)이며 다음 식과 같이 정의된다.

$$\sum_{j=-150}^{150} w_I^2(j) = 1.0 \tag{3}$$

$$r(t) = \sum_{j=-150}^{150} s_{LPF}(j) w_I^2(j) s_{LPF}(j+t) w_I^2(j+t) \tag{4}$$

2.2.1. Look-Back 피치 추정

두 미래 음성 프레임들의 피치들을 P_1와 P_2로 표기하고 과거 두 음성 프레임들의 피치는 P_{-1}과 P_{-2}로 표기한다. 과거의 두 음성 프레임들의 해석을 통해 계산된 초기 피치와 오차 함수를 각각 $\hat{P}_{-1}$, $\hat{P}_{-2}$와 $E_{-1}(P)$, $E_{-2}(P)$로 표기된다. 우선, 오차 함수 $E(P)$는 다음 조건 식 (5)와 식 (6)를 만족시키는 각각 P의 값에 대하여 계산한다.

$$0.8\ \hat{P}_{-1} \le P \le 1.2\ \hat{P}_{-1} \tag{5}$$

$$P \in \{21, 21.5, \cdots, 121.5, 122\} \tag{6}$$

즉 현재의 피치는 과거의 피치값의 20%정도의 범위내에서 변화하도록 하여 음성 피치의 연속성을 보장하는 범위내에서 $E(P)$를 최소로 하는 P의 값을 후방 피치 추정치 (backward pitch estimate) $\hat{P}_B$로 결정한다.

2.2.2. Look-Ahead 피치 추정

미래의 두 음성 프레임들로부터 얻어지는 오차함수를 $E_1(P)$과 $E_2(P)$라 할 때, 미래 프레임들에 대한 피치는 초기에 P_0를 고정시키고 미래의 음성 프레임의 피치 P_1과 P_2를 다음 식 (7), (8), (9)에 의해 $E_1(P) + E_2(P)$를 최소화시키는 $\hat{P}_1$과 $\hat{P}_2$로 결정한다.

$$P_1,\ P_2 \in \{21, 21.5, \cdots, 121.5, 122\} \tag{7}$$

$$8P_0 \le P_1 \le 1.2P_0 \tag{8}$$

$$8P_1 \le P_2 \le 1.2P_1 \tag{9}$$

$\hat{P}_1$과 $\hat{P}_2$로부터 최소 오차함수를 갖는 피치 추정치를 전방 피치 추정치 (forward pitch estimate) $\hat{P}_F$로 결정한다.

일단 후방 피치 추정치와 전방 피치 추정치가 계산되면, 전방 누적 오차 함수와 후방 누적 오차 함수를 비교하여 오차값이 적은 피치 추정치를 초기 피치값으로 결정한다.

2.3. 피치 정제

피치 정제 알고리즘은 피치 추정 정밀도를 1/2 샘플에서 1/4 샘플로 개선하는 과정으로써 주파수 영역에서 새로운 오차 함수를 정의하여 1/4 샘플거리의 예상 피치들에 대한 최소 오차 피치를 추정하게 된다. 열 개의 후보 피치들은 초기 피치 추정값으로 부터 $\hat{P}_I - \frac{9}{8}$, $\hat{P}_I - \frac{7}{8}$,

..., $\hat{P}_I+\frac{7}{8}$ 와 $\hat{P}_I+\frac{9}{8}$ 로 구성한다. 식 (10)에 정의되는 주파수 영역에서의 오차 함수 $E_R(w_0)$은 각각의 피치에 대한 기본 주파수 w_0에 대해서 계산한다. $E_R(w_0)$를 최소로 하는 후보 기본 주파수를 정제된 기본 주파수 $\widehat{w_0}$로 결정하고 이에 따라 피치 값은 $\hat{P}_I=\frac{2\pi}{\widehat{w_0}}$ 로 결정된다.

$$E_R(w_0)=\sum_{m=0}^{\lfloor\left(\frac{9245\pi}{w_0}-0.5\right)\rfloor\frac{256}{2\pi}w_0\rfloor}|S_w(m)-S_w(m,w_0)|^2 \quad (10)$$

여기서 함수 $S_w(m)$은 창함수를 취한 음성신호의 256 샘플 DFT (Discrete Fourier Transform)이다.

III. 피치 예측 방법의 개선

IMBE 음성 부호기의 피치 추정 방법은 1) 모든 음성 프레임을 유성음으로 간주하고, 2) 초기 피치 추정시 현재와 과거의 음성 2 프레임을 이용하는 look-back 피치 추적과, 현재와 미래의 음성 2 프레임을 이용하는 look-ahead 피치 추적을 행하며 3) look-back 피치 추적과 look-ahead 피치 추적에서 구해진 피치들 중 초기 피치로 결정된 피치에 대해서 피치 정제를 통해서 음성의 피치가 결정 된다[1,3,6].

보다 구체적으로 일정한 피치를 갖지 않은 무성음의 경우에도 유성음이나 유/무성음이 혼합된 음성과 동일한 피치 추정을 수행하므로써 불필요한 피치 추정으로 인해 부호기의 복잡도가 증가하게 된다. 또한 초기 피치 추정시 미래의 2 프레임을 필요로 하는 look-ahead 피치 추적 과정으로 인해서 송,수신단에 2 프레임의 시간 지연뿐 아니라 에러함수의 계산 등과 같은 막대한 연산량을 필요로 하여 초기 피치 추정 과정이 전체 부호기 연산 시간의 90% 이상을 차지하게 된다[4,5].

그러므로 본 논문에서는 이와 같은 단점을 보완하기 위해서 1) 초기 피치 추정 전 단계에 유/무성음을 판별하여 유성음으로 결정된 프레임만 피치 추정을 수행하고, 2)음성 신호들의 피치는 아주 느리게 변화하며, 무성음 구간을 제외하고는 대체적으로 일정한 구간에서 변화하는 특징을 고려하여 초기 피치 추정시 미래의 2 프레임을 사용하지 않는 피치 추정 알고리즘을 사용하면 부호기의 연산량이 감소되고 송, 수신단의 2 프레임의 시간 지연이 필요 없게 되어 더욱 효율적인 피치 추정이 가능하게 된다. 그림 2는

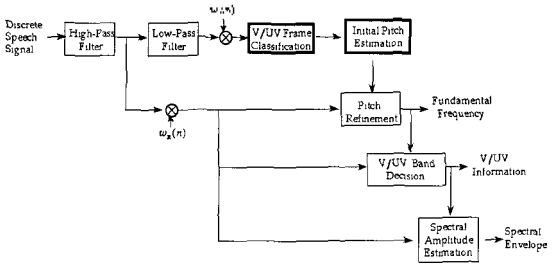

그림 2. 수정된 IMBE 음성 분석 알고리즘
Fig. 2. Modified IMBE speech analysis algorithm.

수정된 IMBE 음성 분석 알고리즘을 나타낸다.

본 논문에서는 초기 피치 추정 전 단계에 유/무성음을 판별하기 위해서 다이애딕 웨이브렛 변환 (Dyadic Wavelet Transform), 영 교차율 (zero crossing rate), 프레임 에너지 (frame energy)를 이용한 알고리즘을 사용하였고, 송,수신단의 시간 지연을 제거하기 위해 미래의 2 프레임을 필요로 하지 않는 개선된 피치 추정 알고리즘을 사용하였다.

3.1. 다이애딕 웨이브렛 변환 (DyWT, Dyadic Wavelet Transform)

다이애딕 웨이블렛 변환 (DyWT)은 연속 웨이브렛 변환에서 스케일 (scale) 변수가 다이애딕 열 (dyadic sequence)로 이산화된 것으로 식 (11)과 같이 정의된다[7-10].

$$DyWT(b,2^j)=\frac{1}{2^j}\int_{-\infty}^{+\infty}x(t)\psi^*\left(\frac{t-b}{2^j}\right)dt \quad (11)$$

여기서 $x(t)$는 입력신호를, $\psi^*(t)$ 는 모함수 웨이블렛 (mother wavelet)의 complex conjugate이며, b는 천이 (translation) 변수이며 2j, j=1, 2, ... 는 스케일 변수이다. DyWT는 음성 신호 분석에 유용한 다음과 같은 특징을 갖는다.

- 선형성과 사불변성: 음성 신호는 주로 천이되고 감쇄되는 사인곡선의 선형 결합으로 모델링되므로 DyWT는 음성 신호 분석에 적합하다.
- 신호의 급격하고 느린 변화의 감지: 신호나 신호의 미분함수가 불연속성을 가지면 신호의 다이애딕 변환의 절대값 $|DyWT(b,2^j)|$는 불연속 점에서 국부 최대값을 갖는다. 따라서 성문의 폐쇄 (glottal closure)와 같은 음성 신호의 급격한 변화를 감지할 수 있다.

피치 주기는 성문이 폐쇄 (glottal closure)되는 순간을

찾아낸 후 성문 폐쇄가 일어나는 시간 간격을 측정함으로써 추정될 수 있다. 성문 폐쇄 시에는 성도 (vocal tract)가 강하게 여기되어서 음성 신호에 급격한 변화를 나타내게 된다. DyWT의 국부 최대값 (local maxima)은 신호의 급격한 변화를 나타내고 국부 최소값 (local minima)은 느린 변화를 나타내므로 음성 신호의 급격한 변화 즉 성문의 폐쇄는 DyWT를 사용해서 감지할 수 있다[9-11].

3.2. DyWT를 이용한 유/무성음 판별 알고리즘

본 논문에서는 DyWT 중 계산량이 적은, 식 (12)로 정의되는 하르 웨이브렛 (Haar Wavelet)을 사용하였다.

$$\phi_H(t) = \begin{cases} 1, & \text{if} \quad 0 < t < \frac{1}{2} \\ -1, & \text{if} \quad \frac{1}{2} < t < 1 \\ 0, & otherwise \end{cases} \tag{12}$$

음성 신호 한 프레임의 DyWT는 각각의 스케일 2j , j = jl, jl+1, … , ju에 대해서 구해지고 b에 대한 국부 최대값을 구한다. 연속적인 두 개의 스케일 2j, 2j+1에 대한 국부 최대값의 위치가 일치하면 이 최대값의 위치에 해당되는 b가 성문 폐쇄가 일어나는 시간에 대응되며, 이 프레임은 유성음 구간으로 판별된다[9,10]. 반면에 연속적인 두 스케일에 대한 국부 최대값의 위치가 일치하지 않으면 성문 폐쇄가 일어나지 않았으므로 무성음 구간으로 판별된다. DyWT는 모든 스케일에 대해서 구해야 하나, 유성음의 기본 주파수는 40Hz ~ 500Hz의 저주파 성분으로 되어 있으며, 무성음은 고주파 성분을 가지므로 스케일 변수의 수를 제한하여 연산량을 줄일 수 있다. 본 논문에서는 j = 1, 2, 3의 세 개의 스케일 값을 사용하여 유/무성음 판별을 수행하였다[11-12].

유/무성음 판별의 정확도를 높이기 위해서 일반적으로 유성음의 경우 높은 에너지를 가지고 있으며, 무성음의 경우 높은 주파수 성분으로 인해 영교차율이 높게 나타나므로, 음성 신호 프레임의 에너지와 영교차율을 동시에 사용하였다.

그림 3은 DyWT를 이용한 유/무성음 판별 알고리즘을 보여주는 블록으로써 다음과 같은 순서로 수행된다.

Step 1: j = 1, 2, 3의 세 개의 스케일에 대해서 DyWT를 구한 후, 연속적인 두 스케일에 대한 국부 최대값의 위치가 일치하는지를 조사한다.

Step 2: 연속적인 두 스케일에 대한 국부 최대값의 위치가 일치하지 않으면 무성음 프레임으로 간주되지만 프레임 에너지가 임계치 (threshold) T3 이상이면 유성음으로 판별한다.

Step 3: 연속적인 스케일에 대한 국부 최대값의 위치가 일치하면 유성음으로 간주되지만 프레임 에너지가 임계치 T1이하면 무성음으로 판별한다. 또한 영교차율이 임계치 T2이상이면 무성음으로 판별한다.

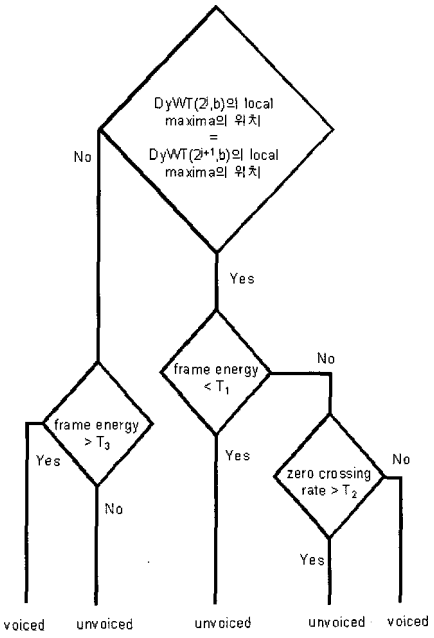

그림 3. DyWT를 이용한 유/무성음 판별 알고리즘
Fig. 3. V/UV decision algorithm using DyWT.

본 논문에서는 T1, T2, T3, 값으로 각각 500, 130, 1000을 사용하였다.

3.3. 개선된 초기 피치 추정 알고리즘

DyWT를 이용해서 유/무성음 판별을 행한 후 무성음으로 결정된 프레임은 특별한 피치가 필요 없기 때문에 피치 추정 과정을 사용하지 않고 랜덤 (random) 피치를 준다. 유성음으로 결정된 프레임에 대해서는 미래의 두 프레임을 사용하지 않고 look-back 피치 추적만을 사용하는 초기 피치 추정을 수행하여 송,수신단의 2 프레임의 시간 지연을 제거하고 초기 피치 추정시 걸리는 처리 시간을 단축한다.

그림 4는 개선된 초기 피치 추정 알고리즘을 나타낸다.

Step 1: 현재 음성 프레임의 $s_{LPF}(n)$에 대해서 식 (12)의 오차함수를 구한다.

Step 2: DyWT를 이용해서 유/무성음 판별을 행한 후 유성음으로 결정되면 look-back 피치 추적에서

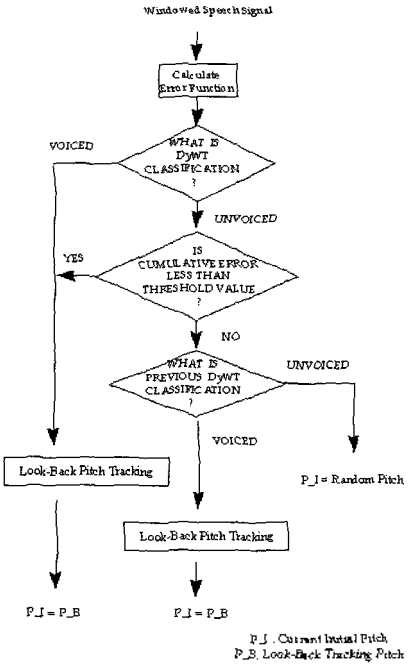

그림 4. 개선된 초기 피치 추정 알고리즘
Fig. 4. Efficient initial pitch estimation algorithm.

얻어진 피치가 초기 피치로 결정된다.

Step 3: 무성음으로 결정된 경우에는 후방 누적 에러 $CE_B(\widehat{P_B})$를 구한 후 후방 누적 에러가 임계치 보다 낮으면 유성음과 같은 방법으로 피치가 구해진다.

Step 4: 후방 누적 에러가 임계치 보다 크면 전 프레임의 유/무성음 여부에 따라서 초기 피치가 구해진다. 전 프레임이 무성음이면 초기 피치로 랜덤 피치를 준다. 반면에 전 프레임이 유성음인 경우에는 피치의 연속성을 고려하기 위해서 look-back 피치 추적에서 얻어진 피치가 초기 피치로 결정된다.

IV. 모의 실험 결과

제안한 부호기의 성능과 음질 평가를 위해서 표 1과 같이 8kHz로 샘플링한 음성 데이터를 사용하였고 펜티엄 150MHz PC를 사용하여 성능 평가를 수행하였다.

피치 추정의 개선이 이루어지지 않은 IMBE 부호기와 피치 추정이 개선된 부호기와의 처리 시간 비교는 표 2와 같다. 표 2의 결과에서처럼 피치 추정 개선이 이루어지지 않은 IMBE 음성 부호기 (IMBEorg, original IMBE)의 프레임당 처리 시간은 340ms이지만 제안한 알고리즘을 사용한 피치 추정이 개선된 IMBE (IMBEpro, proposed IMBE)의 처리 시간은 평균 31.3ms로 처리 시간이 평균 1/11정도 단축되었다.

표 1. 평가에 사용된 음성 데이터
Table 1. Speech data for simulation.

	문 장	프레임 길이	총 프레임수
kf1(한국여성)	미는 피부는 한꺼풀 차이입니다.	20ms	122
km1(한국남성)	이번 겨울은 예년과 달리 포근합니다.	20ms	128
km2(한국남성)	개인통신시대가 조만간에 개막될 것입니다.	20ms	156

표 2. 제안한 IMBE와의 처리 시간 비교
Table 2. Processing time comparison of original and proposed IMBE. [단위: ms/frame]

	IMBEorg	IMBEpro	IMBEorg / IMBEpro
kf1	336	30	11.2
km1	341	33	10.3
km2	344	31	11.1

주) IMBEorg : 피치 추정 개선이 이루어지지 않은 IMBE
IMBEpro : 피치 추정이 개선된 IMBE

표 3. 제안한 IMBE와의 상대적인 처리 시간 비교
Table 3. Relative processing time comparison of original and proposed IMBE.

	Original IMBE (IMBEorg)	Proposed IMBE (IMBEpro)
초기 피치 추정	96.9%	74.0%
피치 정제	2.9%	23.7%
유/무성음 판별	0.1%	1.8%
스펙트럼 진폭 추정	0.1%	0.5%
Total time(ms/frame)	336	30

한 프레임을 처리하는데 걸리는 시간을 100%로 했을 때 피치 추정 개선이 이루어지지 않은 IMBE (IMBEorg)와 제안한 알고리즘을 사용한, 피치 추정이 개선된 IMBE (IMBEpro)의 kf1 (한국여성) 음성 데이터 분석 과정시 각 부분별 상대적인 소요 시간은 표 3과 같다. IMBEorg는 초기 피치 추정이 차지하는 상대적인 소요 시간이 96.9%로 처리 시간의 대부분을 차지하지만, IMBEpro는 초기 피치 추정의 상대적인 소요 시간이 74%로 22.9%가 감소되었음을 알 수 있다.

그림 5, 6, 7은 각 음성 데이터에 대해서 기존의 IMBE와 제안한 IMBE 부호기의 기본 주파수 (피치) 추정 결과를 표시한다. 기본 주파수가 서로 틀린 프레임은 무성음으로 판별된 프레임으로 기존의 IMBE는 피치 추정을 통해

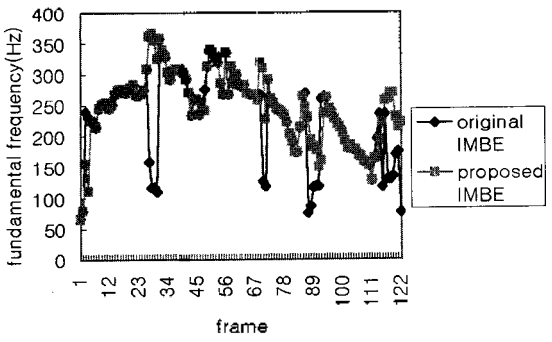

그림 5. kf1 음성 데이터의 기본 주파수 비교
Fig. 5. Fundamental frequency comparison of kf1 data.

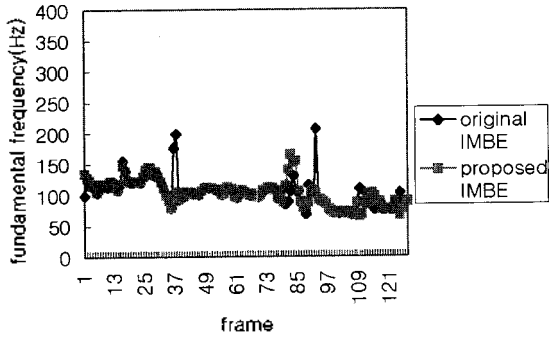

그림 6. km1 음성 데이터의 기본 주파수 비교
Fig. 6. Fundamental frequency comparison of km1 data.

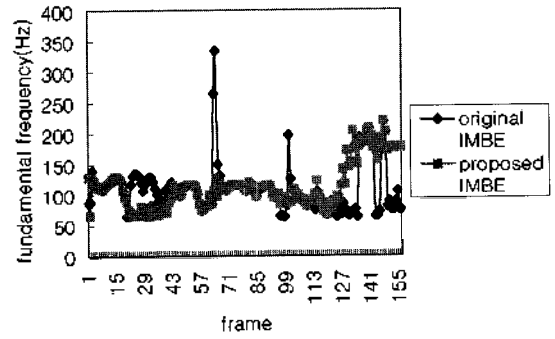

그림 7. km2 음성 데이터의 기본 주파수 비교
Fig. 7. Fundamental frequency comparison of km2 data.

서 기본 주파수를 결정하지만 제안한 부호기는 랜덤 잡음을 기본 주파수 값으로 사용하기 때문에 서로 다른 기본 주파수 값을 갖는다. 그러나 무성음의 경우에는 기본 주파수 값은 의미가 없으므로 복원한 음성 데이터의 음질에는 영향을 주지 않는다. 유성음으로 결정된 프레임의 기본 주파수는 제안한 부호기의 기본 주파수 값과 기존의 IMBE 부호기와 일치함을 알 수 있다.

음성의 질은 여러 사람의 청음을 비교해 본 결과 여성의 음성이 남성보다 우수하였고 기존의 부호화기에 비해 피치 변화가 심한 영역에서는 다소 불규칙한 특성을 보이기도 하였지만 거의 비슷한 음성을 합성하였다.

V. 결 론

기존의 IMBE 음성 부호기의 피치 예측 과정은 초기 피치 예측을 한 후 선택된 후보 피치에 대해서 피치 정제 과정을 거쳐 보다 정밀한 정확도를 갖는 피치를 결정한다. 또한 결정된 피치를 근거로 하여 음성 프레임에 대한 유/무성음 결정과 스펙트럼 진폭의 예측이 이루어지게 된다. 이 과정들 중 초기 피치 추정시 가장 많은 시간이 소요되며 또한 미래의 음성 프레임들이 초기 피치 추정시 사용되므로 시간 지연이 유발되어 실시간 구현에 장애 요소로 작용되었다. 또한 무성음인 프레임에 대해서도 유성음과 동일한 피치 추정을 수행하므로 알고리즘의 효율성을 떨어뜨린다.

본 논문에서는 IMBE 부호기의 문제점을 개선하기 위해서 피치 추정 전 단계에 DyWT, 영교차율, 프레임 에너지를 이용하여 이를 바탕으로 유/무성음 판별을 한 후 유성음으로 결정된 프레임에 대해서만 피치 추정을 행하고 무성음으로 결정된 프레임은 랜덤 잡음을 주어서 부호화시의 처리 시간을 단축하였다. 또한 개선된 피치 추정 알고리즘을 사용하여 피치 추정시 미래의 음성 프레임들을 사용하지 않음으로써 이로 인한 송,수신단의 시간 지연을 줄였다. 그 결과 기존의 IMBE 부호기보다 프레임당 실행 시간이 평균 1/11로 단축되었으며 기존의 부호기와 거의 같은 음질을 얻을 수 있었다.

참 고 문 헌

1. A. M. Kondoz, Digital Speech, John Wiley & Sons, 1996.
2. 이윤근, 안승권, "음성합성기술분야", 전자공학회지, vol. 20, no. 5, May, 1993.
3. D. W. Griffen & J. S. Lim, "Multiband excitation vocoder", *IEEE Trans. on Acoust. Speech and Signal Process.*, Vol. 36, no. 8, pp. 1223~1235, Aug., 1988.
4. 나훈, 윤예섭, 정대권, "웨이브렛 변환을 이용한 IMBE 음성 부호기의 피치 예측 개선에 관한 연구", Proc. of KICS conference 97, vol. 16, no. 2, pp. 1223~1227, 1997.
5. 홍상진, 나훈, 윤예섭, 정대권, "차세대 이동통신용 저속 보코더 개발 및 실시간 구현에 관한 연구", 한국전자통신연구소 최종연구보고서, 1996.
6. Digital Voice Systems, Inc., IMBE vocoder description, 1993.
7. R. K. Young, Wavelet theory and its applications, Kluwer Academic Publishers, 1994.
8. Kenneth R. Castleman, Digital image processing, Prentice-Hall, Inc., 1996.
9. Shubha & G. F. Boudreaux-Bartels, "A comparison of a wavelet functions for pitch detection of speech signals", *Proc. of ICASSP*, pp. 449~452, May 1991.

10. Shubha & G.F.Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals" , *IEEE Trans. on Information theory*, vol. 38, no.2, March 1992.
11. L.R. Rabiner & R. W. Schafer, Digital processing of speech signals, Prentice-Hall, Inc., 1978.
12. 정재호, "3.8kbps 호모몰픽 보코더" , Proc. of KSPC conference 92, vol. 5, no. 1, pp.19~22, 1992.

저자 약력

• 나 훈 (Hoon Na)

1996년: 한국항공대학교 항공전자공학과 (공학사)
1998년: 한국항공대학교 항공전자공학과대학원 (공학석사)
1999년~현재: 한국항공대학교 항공전자공학과대학원 박사과정 재학
※ 주관심분야: 정지영상 및 동영상 압축기법, 멀티미디어 신호처리, 음성 부호기, VoIP

• 정 대 권 (Dae-Gwon Jeong)

1979년: 한국항공대학교 항공전자공학과 (공학사)
1987년: Texas A&M대 대학원 (공학석사)
1990년: Texas A&M대 대학원 (공학박사)
1979년~1984년: 국방과학연구소 연구원
1990년~1991년: 한국전자통신연구원 선임연구원
1991년~현재: 한국항공대학교 항공전자공학과 교수
※ 주관심분야: 정지영상 및 동영상 압축기법, 저속음성 부호기법, IMT-2000 부호기, JPEG2000 부호기, 멀티미디어 신호처리