2.9kbps LP-SMBE 음성부호기 개발

박 승 종** , 김 승 주* , 오 영 환* , 이 양 희**
*한국과학기술원 전산학과 , **동덕여자대학교 전자계산학과

Design of the 2.9kbps LP-SMBE vocoder

Seung-Jong Park**, Sung-Joo Kim* , Yung-Hwan Oh* , Yang-Hee Lee**
*Dept. of Computer Science KAIST , **Dongduck Women's Univ.

요 약

본 논문에서는 선형 예측 방법(linear prediction : LP)과 다중 대역 여기 방법(multi-band excitation : MBE)의 장점을 조합하여 낮은 전송률에서 고품질의 합성음을 제공하는 LP-SMBE(linear prediction - simplified multi-band excitation) 부호기를 제안한다. LP-SMBE 부호기에서는 선형 예측 방법과 단순화된 여기 신호 추정 방법을 이용하여 성도 특성 정보와 여기 신호를 분리 추정한다. 제안한 단순화된 여기 신호 추정 방법은 정규화된 스펙트럼 영역에서 원음 스펙트럼과 합성 스펙트럼을 비교하여 여기 신호를 추정한다. 이 방법은 기존 MBE방법의 여기 신호 추정 방법보다 연산량이 적고, 여기 신호를 보다 정확히 추정할 수 있다.

제안한 방법을 이용하여 전송률 2.9kbps 음성부호기를 개발하였다. 개발된 부호기는 연산량을 4.8kbps DoD CELP 연산량의 1/3, 4.8kbps VSELP 연산량의 1/2로 감소시켰다. 또한 합성 음질 선호도 실험에서도 다른 두 부호기의 합성 음질보다 뛰어난 결과를 나타내었다.

1. 서 론

음성부호화 기술은 음성 신호의 중복성(redundancy)을 제거하여 정보량을 줄이는 기술로서, 음성 전송시 전송선의 효율을 높여주며, 저장시 기억용량을 줄여준다. 음성부호화 기술은 크게 파형 부호화 방법(waveform coding)과 음원 부호화 방법(source coding), 그리고 두 방법의 장점을 혼합한 복합형 부호화 방법으로 분류할 수 있다. 이중에서 8kbps 이하의 낮은 전송률에서 우수한 합성음질을 제공하는 음성 부호화 방법으로서 음원 부호화 방법과 복합형 부호화 방법이 많이 연구되고 있다. 대표적인 음원 부호화 방법으로는 다중 대역 여기 부호화 방법이 있고, 복합형 부호화 방법으로는 벡터 여기(vector excitation) 부호화 방법이 있다.

벡터 여기 부호화 방법은 여기 신호를 벡터 양자화(vector quantization)하는 파형 부호화 방법과, 음성의 조음 특성을 반영하는 음성 합성 필터를 선형 예측 방법(linear predictive analysis)을 사용하여 추정하는 음원 부호화 방법을 혼합하여 사용한다. 현재 연구되고 있는 대표적인 벡터 여기 부호화 방법으로는 CELP(code excited linear prediction), QCELP, 그리고 VSELP(vector sum excited linear prediction) 부호기등이 있다[1]. 이러한 부호화 방법은 4.8kbps 이상의 전송률에서는 고품질의 합성음을 제공하지만, 4kbps 이하의 낮은 전송률에서는 합성음질이 급격히 저하되며, 많은 연산량을 필요로 하는 문제점이 있다[8].

다중 대역 여기 부호화 방법은 기존의 음성 생성 모델들이 음성구간 전체를 유성음 혹은 무성음으로 구분하는 것과는 달리, 음성구간의 주파수 영역을 여러개의 대역으로 분할한 후, 각 대역에 대하여 유/무성음을 구분하는 다중 대역 여기 음성 생성 모델을 사용한다. 현재 연구되고 있는 부호기로는 IMBE (improved multi-band excitation)가 있으며, 4kbps 이하의 전송률에서도 고품질의 합성음을 제공하며, 벡터 여기 부호화 방법보다 연산량이 적은 것으로 알려졌다[5]. 그러나 IMBE 부호기도 2.4kbps 이하의 낮은 전송률을 유지하기 위해서는 음원 파라미터들을 양자화할 때 벡터 양자화 방법등을 사용하므로 많은 연산량을 필요로하는 문제점이 있다[6].

본 논문에서는 적은 연산량으로 4kbps 이하의 낮은 전송률을 유지하기 위하여 음원 정보와 성도 특성 정보를 분리하여 추정하는 선형 예측 - 단순화된 다중 대역 여기 방법을 제안한다. 제안한 부호기에서는 4kbps 이하의 낮은 전송률을 유지하면서 벡터양자화보다 연산량이 적은 선형 예측 방법을 사용하여 성도 특성 정보인 스펙트럼 포락을 추정한다. 또한 적은 연산량으로 고품질 합성음을 제공하는 단순화된 다중 대역 여기 방법을 사용하여 음원 정보인 피치와 유/무성음 정보를 추정한다.

2. LP-SMBE 부호기

2.1 LP-SMBE 음성 생성 모델

LP-SMBE 음성 생성 모델은 식 (1)과 같이 선형 예측 방법을 사용하여 성도 특성 정보 $H_{LPC}(\omega)$를 표현하며, 단순화된 다중 대역 여기 방법을 사용하여 여기 신호 $|E_{SMBE}(\omega)|$를 표현한다. 제안한 음성 생성 모델은 성도 특성 정보와 여기 신호를 분리 추출함으로써 정보량을 줄일 수 있으며, 여기 신호의 추정 방법을 단순화시킬 수 있다.

$$\hat{S}(\omega) = H_{LPC}(\omega)|E_{SMBE}(\omega)| \quad (1)$$

성도 특성 정보를 표현하는 방법에는 여러가지가 있다. 기존의 다중 대역 여기 부호기에서는 스펙트럼 상의 각 고조파에서 계산된 스펙트럼 진폭을 사용하여 성도 특성 정보를 표현한다. 이 방법은 스펙트럼 포락을 정확히 묘사할 수 있으나 정보량이 많아 양자화시 문제가 된다. 그러므로 본 논문에서는 정확성에서는 떨어지지만 정보량이 적은 선형 예측 계수를 사용한다. 실험적으로 선형 예측 계수를 사용하여 MBE 모델의 스펙트럼 포락을 표현할 경우, 합성음의 음질이 저하되지 않는 것으로 나타났다.

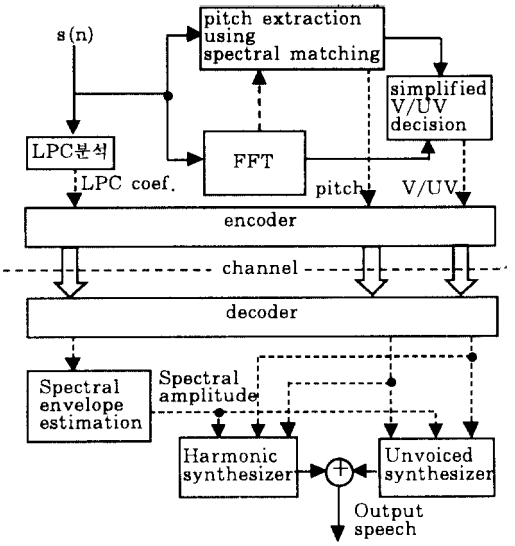
여기 신호 표현 방법 중에서 MBE 모델은 다른 여기 모델에 비하여 적은 정보량으로 우수한 합성음질을 제공한다. 본 논문에서는 MBE 모델을 근간으로 적은 연산량으로 MBE 모델의 파라미터를 효율적으로 추정하는 단순화된 다중 대역 여기 추정 방법을 제안한다. 기존의 MBE 모델에서는 스펙트럼 영역에서 합성에 의한 분석 방법을 사용하여 원음 스펙트럼과 합성음 스펙트럼의 오차가 최소화 되도록 여기 신호를 추정한다. 그러나 제안한 단순화된 MBE 모델에서는 정규화된 스펙트럼 영역에서 피치와 각 대역별 유/무성음 구분 정보를 추정한다. 제안된 추정 방법은 기존의 방법보다 적은 연산량으로 여기 신호를 정확히 추출할 수 있다.

2.2 분석부

제안한 부호기에서 추정하는 음원 파라미터는 스펙트럼 포락과 피치, 그리고 유/무성음 정보이다.

기존 MBE 부호기에서는 합성에 의한 분석법을 사용하여 스펙트럼 포락과 여기 신호를 동시에 추정하게 된다. 동시에 추정한 파라미터 중에서 스펙트럼 포락은 각 고조파 대역별로 샘플링되므로 정보량이 많은 단점이 있다. 위의 단점을 극복하기 위하여 2.4kbps IMBE 부호기는 벡터 양자화 방법을 이용하여 스펙트럼 포락을 양자화한다[5]. 그러나 벡터 양자화 방법을 이용할 경우 연산량이 많아지는 문제점이 발생한다.

위의 문제점을 해결하기 위해서 본 논문에서는 스펙트럼 포락과 여기 신호를 분리하여 추정하는 방법을 제안한다. 스펙트럼 포락은 연산량이 적으면서 양자화시 정보량이 적은 선형 예측 방법을 이용하여 추정하고, 여기 신호의 추정은 단순화된 여기신호 추정 방법을 이용한다. 단순화된 여기신호 추정 방법이란 각 고조파 대역별로 정규화된 스펙트럼과 음성 스펙트럼을 비교하여 피치와 각 대역별 유/무성음을 추정하는 방법이다.



(그림 1) 제안한 부호기 구조도

1) 성도 특성 정보 추정

제안된 부호기의 분석부는 스펙트럼 포락을 추정하기 위해서 선형 예측 방법을 이용하여 입력 음성으로부터 10차 선형 예측 계수를 자기 상관 방법을 사용하여 구한다. 선형 예측 계수는 전송시 양자화 특성이 좋은 LSP(line spectrum pair)계수로 변환하여 전송한다.

2) 피치 주기 추정

다중 대역 여기 부호 방법으로 여기 신호를 추정하는 부호기는 피치 오류가 발생할 경우, 합성음질이 급격히 저하되는 단점이 있다. 이는 MBE 부호기가 추정된 피치를 사용하여 대역을 분할하므로 대역이 잘못 분할될 경우, 각 고조파 대역에서의 유/무성음 정보를 잘못 추정하기 때문이다. 피치 탐색 오류 중에서 가장 많이 발생하는 오류는 정수배 피치 선택 오류와 역정수배 피치 선택 오류가 있다.

기존 MBE 방식의 부호기들은 피치 탐색 오류를 제거하기 위하여 동적 프로그래밍 방법을 이용한 피치 추적(pitch tracking)을 수행하여 피치 탐색 오류를 줄이고 있다. 그러나 위의 피치 추적 방법은 현재 음성구간에 인접한 앞뒤 음성 구간들의 피치를 동시에 고려하므로 음성 부호화시 많은 지연 시간이 발생한다. 또한 피치탐색 알고리즘의 복잡도가 매우 크다.

본 논문에서는 부호화시 지연 시간이 적고 알고리즘이 단순하며 연산량도 적은 스펙트럼상에서의 피치 탐색 방법을 제안한다.

단계 1. 후보 피치 선정

자기 상관 방법을 이용하여 K(= 4)개 후보 피치를 선정한다. K개 후보 피치들에는 원음의 피치와 정수배 피치(multiple pitch)들, 그리고 역정수배 피치(submultiple pitch)들이 포함된다.

단계 2. 정규화된 합성 스펙트럼 생성

단계 1.에서 선정된 후보 피치들을 이용하여 K개의 주기적 스펙트럼을 합성한다. 합성된 주기적 스펙트럼의 첫번째 고조파 대역의 진폭을 해당 대역에서의 최대 스펙트럼 진폭으로 나누어 0에서 1의 값을 갖도록 정규화한다.

단계 3. 정수 피치 추정

K개의 정규화된 주기적 스펙트럼 $E(\omega)$ 중에서 식 (2)의 오차 ε를 최소화시키는 스펙트럼을 선택한다. 선택한 주기적 스펙트럼의 피치를 최종적인 정수피치로 선택한다.

$$\varepsilon = \frac{1}{b_1 - a_1} \int_{a_1}^{b_1} \left| \tilde{S}(\omega) \cdot E(\omega) \right| d\omega \qquad (2)$$

식 (2)에서 a_1, b_1은 첫번째 고조파 대역의 저역 주파수와 고역 주파수이며, $\tilde{S}(\omega)$은 원음 스펙트럼을 첫번째 고조파 대역에서 최대값으로 나누어 구한 정규화된 스펙트럼이다.

각 단계들을 거쳐 선택된 정수 피치를 원음과 합성 스펙트럼의 오차를 최소화 하는 비정수 피치로 정교화한다[2].

3) 유/무성음 정보 추정

여기 신호인 유/무성음 결정은 음성 구간의 피치에 따라 10에서 50개의 고조파 대역에 대하여 이루어 진다. 각 고조파 대역의 유/무성음 정보는 정규화된 스펙트럼 영역에서 스펙트럼의 비교를 이용하여 추정한다.

유/무성음 정보를 추정하기 위하여 먼저, 추정한 피치를 이용하여 주기적 스펙트럼 $E(\omega)$를 합성한다. 합성된 주기적 스펙트럼은 0에서 1의 값을 갖도록 정규화된 스펙트럼이다. 다음으로 원음 스펙트럼을 각 고조파 대역별로 정규화한다. 정규화는 각 고조파 대역에서의 진폭을 해당 대역에서의 최대 진폭으로 나누어 0에서 1의 값을 갖도록 한다. 정규화된 원음 스펙트럼 $\tilde{S}(\omega)$와 주기적 스펙트럼 $E(\omega)$을 비교하여 식 (3)의 오차 ε_m를 계산한다.

$$\varepsilon_m = \frac{1}{b_m - a_m} \int_{a_m}^{b_m} \left| \tilde{S}(\omega) \cdot E(\omega) \right| d\omega \qquad (3)$$

계산한 오차 ε_m가 임계치를 초과할 경우 해당 고조파 대역에서는 주기적 스펙트럼에서 발생하는 고조파 스펙트럼이 발생하지 않으므로 무성음으로 결정한다. 반대로 임계치보다 작을 경우 유성음으로 결정한다.

2.3 합성부

제안한 부호기의 합성부에서는 LPC계수로부터 스펙트럼 포락을 계산한 후, 스펙트럼 포락, 피치, 유/무성음 정보들을 사용하여 유성음과 무성음을 각각 합성한다. 합성과정은 다음과 같다.

1) 스펙트럼 포락 계산

전송된 에너지 G와 LPC계수의 임펄스 응답(impulse response)을 푸리에 변환하여 구한 $|A(e^{j\omega})|$를 식 (4)와 같이 계산하여 스펙트럼 포락 $|H(\omega)|$을 구한다[7].

$$|H(\omega)| = \frac{G}{\left| 1 + \sum_{k=1}^{p} a_k e^{-jk\omega} \right|} = \frac{G}{|A(e^{j\omega})|} \qquad (4)$$

2) 유성음 합성

유성음은 시간 영역에서 각 고조파에 해당하는 삼각 함수들의 합으로써 식 (5)와 같이 합성한다.

$$\hat{S}_v(t) = \sum_{m=1}^{L} A_m(t) \cos(\theta_m(t)) \qquad (5)$$

식 (5)에서 진폭함수 $A_m(t)$는 분리된 유성음 스펙트럼을 프레임간에 선형보간한 함수이다. 이 때 무성음으로 판정된 고조파의 진폭 A_m 은 영으로 가정한다. 그리고 위상함수 $\theta_m(t)$ 는 초기위상과 기본 주파수의 함수를 사용하여 계산한다.

3) 무성음 합성

스펙트럼 영역에서 스펙트럼 포락과 여기 파라미터로부터 합성 스펙트럼을 구한 후, 역푸리에 변환을 사용하여 무성음을 생성한다.

3. 실험 및 평가

3.1 실험 환경

본 논문에서는 단순화된 여기 신호 추정 방법과 선형 예측 방법을 조합하여 2.9kbps LP-SMBE 부호기를 개발하고, 성능을 평가 하였다. 개발된 LP-SMBE 부호기의 분석구간(frame)의 길이는 25ms이고 분석구간의 이동 간격은 20ms 이다. 부호기의 전송 파라미터에 할당된 비트수는 [표 1]에 보였다.

음질 평가는 주관적 선호도(preference test)를 4.8kbps DoD CELP와 4.8kbps VSELP에 대하여 각각 비교 실험하였다. 비교 실험에 참가한 인원은 20명이며, 음성 자료는 8 kHz 표본 추출된, 총 32초 길이의 4개 한국어 문장을 사용하였다.

[표 1] 제안한 부호기의 bit 할당

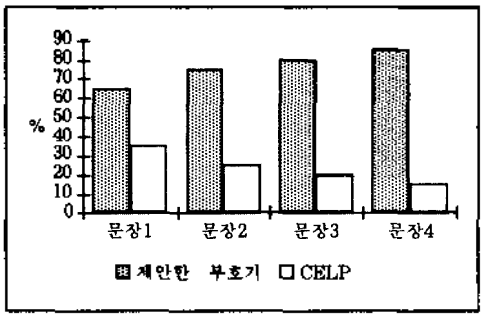
	bits/frame	bits/sec
LSP	34 bits	1700 bits
energy	6 bits	300 bits
pitch	8 bits	400 bits
V/UV	10 bits	500 bits
Total	58 bits	2900 bits

〔표 2〕 음성 자료

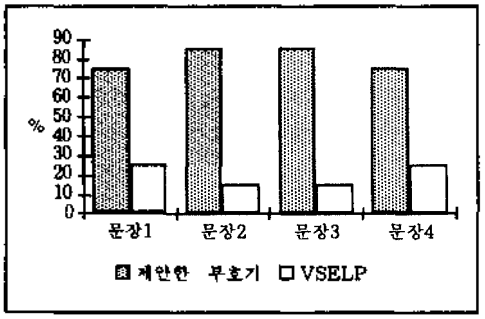
문장 1	남성 1인	clean speech	9.19초
문장 2	여성 1인	clean speech	8.17초
문장 3	남성 1인	clean speech	4.80초
문장 4	남성 1인	noisy speech	9.19초

3.2 실험 결과

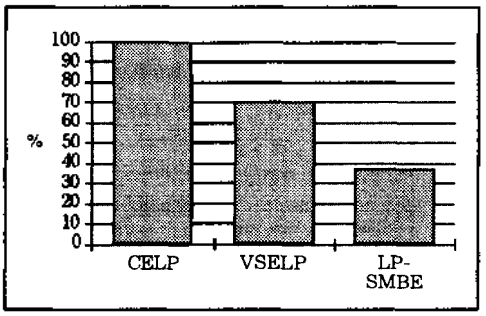
합성 음질을 비교 평가한 결과 잡음이 없는 음성과 잡음이 섞인 음성들에 대해서 제안한 LP-SMBE 부호기의 합성 음질이 CELP와 VSELP부호기의 합성 음질보다 뛰어난 것으로 나타났다. (그림 2)와 〔그림 3〕은 제안한 부호기의 합성음질을 CELP와 VSELP에 대하여 선호도 실험을 수행하여, 선호도를 백분률로 환산하여 나타낸 그림이다.



(그림 2) 제안한 부호기와 CELP와의 합성음질 비교



〔그림 3〕 제안한 부호기와 VSELP와의 합성음질 비교



〔그림 4〕 연산량 비교

제안한 부호기와 다른 두 부호기의 연산량을 비교하기 위해서 소프트웨어 모의 실험에서의 처리속도를 비교하였다. 실험결과, 제안된 부호기는 CELP 연산량의 1/3, VSELP 연산량의 1/2에 해당하는 연산량을 갖는 것으로 나타났다. 〔그림 4〕는 CELP의 연산량을 기준으로 LP-SMBE와 VSELP의 연산량을 백분률로 환산한 그림이다.

4. 결론

본 논문에서는 선형 예측 방법과 단순화된 다중 대역 여기 방법을 사용하여, 성도 특성 정보와 여기 신호를 분리하여 추정하는 2.9kbps LP-SMBE 부호기를 제안한다. 선형 예측 방법을 사용한 성도 특성 정보 추정 방법은 전송률이 낮으면서, 동시에 단순화된 여기 신호 추정 방법을 사용할 수 있게 한다. 단순화된 여기 신호 추정 방법은 기존의 다중 대역 여기 신호 추정 방법보다 적은 연산량으로 여기 신호를 보다 정확히 추정한다.

제안된 부호기를 4.8kbps DoD CELP와 4.8kbps VSELP 부호기에 대해서 각각 비교 실험을 수행하였다. 합성 음질 선호도 실험에서는 다른 두 부호기보다 제안된 부호기의 합성 음질이 우수하며, 연산량은 2배 이상 감소한 것으로 나타났다.

향후 연구과제로는 삼각함수를 사용함으로써 발생하는 합성부의 많은 연산량을 줄이기 위한 연구가 필요하다고 판단된다.

참고 문헌

[1] I. A. Gerson, and m. A. Jasiuk, "Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8kbps", Proc. of ICASSP'90, pp.461-464, 1990.

[2] Daniel W. Griffin and Jae S. Lim, "Multi-Band Excitation vocoder," IEEE Trans. on Acoustics, Speech and Signal Proc., vol. ASSP-36, pp.1223-1235, Aug. 1988.

[3] Michael S. Brandstein, Peter A. Monta, John C. Hardwick, and Jae S. Lim, "A Real-time Implementation of the Improved MBE Speech Coder," Proc. of ICASSP'90, Albuquerque, NM, April 3-6, 1990.

[4] Siu-Wah Wong, "An Evaluation of 6.4kbit/s Speech Codecs for INMARSAT-M System," Proc. of ICASSP'91, pp.629-632, 1991

[5] P. C. Meuse, "A 2400 bps Multi-Band Excitation Vocoder", Proc. of ICASSP'90, pp.9-12, 1990.

[6] Masayuki Nishiguchi, Jun Matsumoto, Ryoji Wakatsuki, and Shinobu Ono, "Vector Quantized MBE with Simplified V/UV Decision at 3.0kbps," Proc. of ICASSP'93, pp.151-154, 1993.

[7] J. D. Markel, A.H. Gray, Jr., Linear Prediction of Speech, 1976.

[8] S. J. Kim, S. J. Park, Y. H. Oh,"Complexity Reduction Methods for Vector Sum Excited Linear Prediction Coding," International Conference on Spoken Language Processing, pp.2071-2074, 1994.