단순화된 다중 모드 방법을 이용한 음성 부호화기

강 홍 구*, 이 인 성**, 차 일 환*, 윤 대 희*

* 연세대학교 전자공학과
** 한국전자통신연구소 신호처리 연구실

A Speech Coder using the Simplified Multi-mode Method

H. G. Kang*, I. S. Lee**, I. W. Cha*, D. H. Youn*

* Department of Electronic Engineering Yonsei University
** Electronics and Telecommunications Research Institute

ABSTRACT

This paper proposes a SM-CELP(Simplified Multi- mode CELP) speech coder which applies different excitation signal according to the characteristic of speech segment at bit-rates below 4 kbps. Speech signal is divided with 2 modes such as stationary voice(mode-A) and etc.(mode-B) using the parameters of average energy of the short-time speech and the residual signal after long term prediction.

Structured multi-pulse method is used for the excitation of mode-A and gaussian or pulse-like codebook for mode-B. 4.8kbps DoD-CELP are used to evaluate the performance of the proposed coder. As a result, the proposed method shows 1~2 dB higher segmental signal to noise ratio and better subjectional quality without increasing the computational amount.

I. 서 론

음성 관련 분야 중 신호를 압축한 후 왜곡없이 복원하기 위한 부호화 연구는 디지탈 음성 통신 분야에서 매우 중요한 역할을 담당하고 있다. 이러한 연구는 디지탈 셀룰라, 음성 저장 및 전송 시스템 등을 중심으로 활발히 연구되고 있으며, 선형 예측(linear prediction)을 기반으로 하는 분석-합성(analysis-by-synthesis) 방법이 주로 사용되고 있다[1].

최근에는 4kbps 내외의 전송률을 갖는 half-rate 에서도 기존 시스템과 비슷한 성능을 유지하기 위한 연구를 통해 낮은 전송률 음성 부호화 기술은 급격한 발전을 이루고 있으며, 이를 바탕으로 각 국에서는 표준안을 정하기 위한 작업이 진행중이다[2].

기존의 분석-합성 부호화 방법과 같은 구조하에서 전송률을 낮추고 만족할 만한 성능을 유지하기 위한 한 방법은 일정 변수를 통해 분석 구간의 특성을 미리 파악한 후, 각각에 적합한 여기 신호를 사용하는 다중 모드(multi-mode)를 이용하는 것이다[4][5]. 대표적인 다중 모드 방법인 M-LCELP(Multi-mode Learned CELP)[4]는 북미 half-rate 디지탈 셀룰라 표준화를 위해 제안된 방법으로서 그림 1과 같은 분석-합성 구조를 지닌다.

본 연구는 한국전자통신연구소 연구비 지원에 의한 결과임.

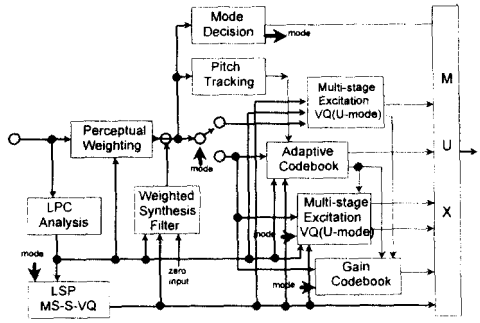

그림 1. M-LCELP 블럭도

모드 결정 블럭에서는 3개의 임계값으로 구성된 피치 예측 이득을 이용하여 모드-0, 모드-1, 모드-2, 모드-3의 네 가지로 구분하며, 결정된 모드 정보에 따라 각 구간의 여기 신호 생성 방법이 달라진다. 모드-1, 2, 3에서는 간축비를 2로 갖는 일정 간격의 펄스(regular pulse) 코드북을 2단으로 그리고, 무성음으로 이루어지는 모드-0에서는 3단으로 구성된 비정기적 펄스 코드북을 사용하여 여기 신호를 생성한다. 코드북을 다단으로 구성함에 따른 계산량과 성능과의 trade-off 관계에 의해 첫번째 단에서는 몇 개의 후보를 미리 선택하고, 두번째 단에서는 선택된 코드와의 조합을 통해 최적의 코드를 찾는 방법을 사용한다.

M-LCELP에서 위와 같이 여기 신호를 복잡하게 생성하는 이유는 전송률을 4 kbps로 낮추기 위해 프레임 길이를 40 msec로 확장하였기 때문이다. 프레임의 길이가 이처럼 길어지면 음성의 구간 특성 변화가 심해지므로 스펙트럼 포락선 성분을 정확히 분리하기 어려워진다. 따라서, 구분한 모드 각각에 대해서도 여기 신호를 좀 더 세밀하고, 정교하게 처리해야만 좋은 음질을 유지할 수 있다. 또한, 이 방법은 프레임 길이가 길어짐에 따라 부호화 지연 시간(coding delay)이 길어지므로 실제 시스템에 적용하기 어렵다는 단점도 발생한다.

본 논문에서는 이러한 문제를 극복하기 위해 프레임의 길이를 축소하고, 모드 변화에 따라 효율적인 여기 신호를 적용함으로써 성능 개선을 시도하였다. 논문의 구성은 다음과 같다. 2 장에서는 제안된 부호화기의 모드를 구분하기 위한 알고리듬과 각 모드에 따른 여기 신호 생성 방법을 서술한다. 3 장에

서는 4kbps 이하의 평균 전송률을 유지하기 위해 각 변수에 할당된 비트 정보와 이를 이용하여 평균 3.6 kbps로 구현된 부호화기의 주/객관적 성능 평가에 대해 알아본다. 성능 평가는 표준안으로 제정된 방법과의 비교가 객관적으로 타당하므로 4.8kbps DoD-CELP(Department of Defense CELP)[6]를 사용한다. 마지막으로 4장에서 결론을 맺는다.

II. 단순화된 다중 모드 방법

제안한 단순화된 다중 모드 방법의 블럭도는 그림 2와 같다.

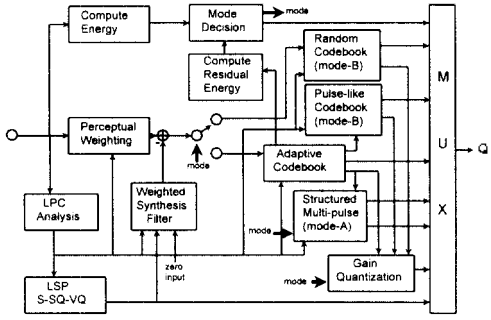

그림 2. 단순화된 다중 모드 부호화기의 블럭도

우선 프레임 에너지와 잔차 신호 에너지를 이용하는 간단한 구분 알고리듬을 통해 음성을 정상 상태의 유성음(모드-A)과 그 외 구간(모드-B)으로 구분한다. 유성음의 특징이 가장 잘 나타나는 모드-A 구간에서는 구조화된 다중 펄스열[7]을 사용하여 명료성을 높이고, 연산량에서의 이득을 취한다. 특성이 변하는 유성음 구간과 무성음 구간으로 구분되는 모드-B에서는 이전 프레임의 모드에 따라 가우시안 백색 잡음 혹은 펄스 형태로 구성된 코드북을 사용한다.

2-1. 모드 구분법

다중 모드 방법에서는 구간 신호의 특성을 구분하기 위한 전처리 단계가 반드시 필요하다. 기존 방법에서는 분석 프레임의 길이가 길어짐에 따라 복잡한 변수를 통해 모드를 구분했지만, CELP 알고리듬 특성 상 모드 구분을 정교하게 하여도 큰 이득을 얻기 어렵다. 따라서, 본 논문에서는 프레임의 길이를 줄인 후, 단구간 입력 신호 에너지와 잔차 신호 에너지와 같이 단순하면서도 구간의 특징을 비교적 효율적으로 반영할 수 있는 변수를 이용한다.

가. 단구간 에너지

입력된 일정 구간의 음성 신호는 (1) 식에 주어진 n 번째 프레임에서의 단구간 에너지를 이용하여 우선 유성음 구간과 그 외 구간으로 구분한다.

$$E_s(n)=\sum_{m=-\infty}^{\infty} s^2(m)w(n-m) \qquad (1)$$

여기서, $s(m)$은 입력 음성 신호이며, $w(n)$은 창(window) 함수이다. 실험을 통해 임계값(THD_1)을 구한 후, 단구간 에너지가 임계값보다 작은 경우에는 모드-B 로 판별하며, 클 경우에는 다음 절에 주어진 단구간 잔차 신호 에너지를 이용하여 다시 세분한다.

$$E_s(n)=\begin{cases} > |THD_1|, & \text{유성음 구간} \\ < |THD_1|, & \text{무성음/묵음 구간 : 모드}-B \end{cases} \qquad (2)$$

나. 잔차 신호 에너지

(2) 식에서 유성음으로 판별된 구간에 대해서는 다시 잔차 신호의 에너지를 계산하여 장구간 예측 성능을 조사한다. 특성이 변하는 유성음 부분의 잔차 신호에는 펄스형 성분이 복잡하게 존재하므로 평형 상태의 유성음에 비해 에너지가 크다. 이런 구간에서는 비슷한 크기를 갖는 펄스 성분이 다수 존재하므로 여기 신호 발생시 무성음 구간과 같이 취급하는 것이 효율적이다.

이를 구분하기 위해 (3) 식과 같이 정규화된 잔차 신호의 에너지를 사용한다.

$$E_r(n)=\sum_{m=-\infty}^{\infty} r^2(m)w(n-m) \qquad (3)$$

여기서 $r(n)$은 장구간 예측 후의 신호이며, $w(n)$은 창(window) 함수이고, N 은 부프레임 길이이다. 단구간 에너지 판별법에서와 마찬가지로 임계값(THD_2)은 실험을 통해 얻는다.

$$E_r(n)=\begin{cases} > |THD_2|, & \text{평형 상태 유성음 : 모드}-A \\ < |THD_2|, & \text{특성이 변하는 유성음 : 모드}-B \end{cases} \qquad (4)$$

구간 에너지와 잔차 신호의 에너지로 부터 무성음 구간과 특성이 변하는 유성음 구간은 동일한 모드로 구분되므로 결국 모드는 두 종류로 구성된다. 그림 3은 음성 신호를 사용하여 실제로 모드를 구분한 예를 나타낸다.

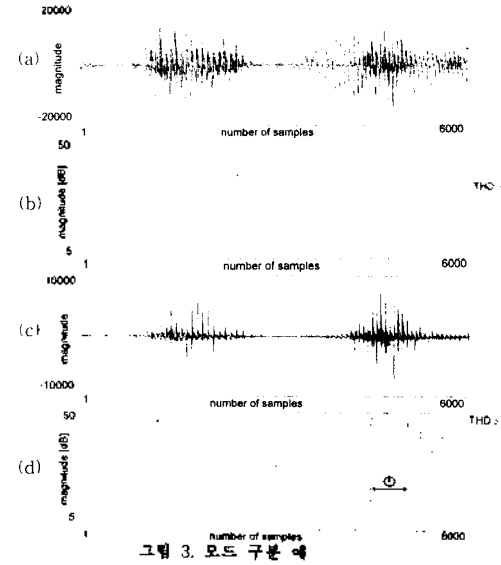

그림 3. 모드 구분 예

(a) 원음
(b) 단구간 에너지($E_s(n)$)
(c) 잔차 신호
(d) 잔차 신호 에너지($E_r(n)$)

그림 3에서 (a)는 원음 (b)는 (a)로 부터 프레임마다 구한 에너지이고, (c)는 잔차 신호이며, (d)는 잔차 신호의 에너지를 부프레임 단위로 구한 값이다. (b)에서 점선으로 표시된 THD1 보다 작은 값을 갖는 영역은 모드-B로 선택되며, 큰 값을 갖는 영역에 대해서는 부프레임에 대해 다시 한번 에너지를 구한다 (d)에 주어진 임계값 THD2는 특성이 변하는 유성음 구간을 판단하기 위해 주어진 값이다. ↓로 표시된 곳은 특성이 변하는 유성음 구간을 나타내는 부분으로서 본 논문에서는 모드-B로 판단하는 부분이다.

2-2. 모드 결정에 따른 부호화 방법

합성음의 성능을 높이기 위해서는 모드 구분에 따라 적합한 여기 신호를 발생하는 것이 중요하다. 유성음 구간에서는 잔차 신호에 존재하는 피치 성분을 모델링하는 것이 효과적이며, 무성음 구간에서는 백색 잡음 특성을 반영하는 것이 효과적이다. 본 논문에서는 낮은 전송률에서도 이러한 조건을 만족할 수 있도록 하는 여기 신호 생성 방법을 제안하고, 고속 탐색 알고리듬을 통해 연산량을 감소시킬 수 있는 방법을 제안한다. 또한, 이전 구간의 모드 정보를 이용하여 모드 변화에 따라 추가 정보 없이도 여기 신호를 효과적을 응용할 수 있는 방법을 제안한다

가. 명청 상태 유성음 : 모드 A

유성음 구간에서의 명료성을 높이기 위해서는 피치 성분을 정확히 모델링하는 것이 중요하다. 이를 위해 그림 4와 같이 구조화된 펄스 조합으로부터 최적의 펄스 위치와 간격을 결정하는 구조화된 다중 펄스 방법을 사용한다[7].

구조화된 다중 펄스 방법은 MPE(Multi Pulse Excitation)[8]에서와 같이 다중 펄스열을 도입하여 장구간 예측 후의 신호에 존재하는 펄스 성분을 모델링하는 방법으로서 계산량 및 전송률을 줄이기 위해 펄스 갯수와 탐색 영역, 그리고 펄스 양자화 레벨을 몇 개로 제한하는 방법이다[7]. 이를 자세히 설명하면 다음과 같다.

제한조건-1 : 부프레임 당 펄스 갯수는 전송률 및 연산량을 고려하여 고정시킨다.

제한조건-2 : 첫번째 펄스의 크기는 "1"이며 부프레임의 어느 위치에나 존재할 수 있다.

제한조건-3 : 두번째 이후의 펄스 크기는 $\pm 2^{-n}$ 으로 양자화하고, 이전 펄스부터의 일정 간격에 대해 완전 탐색(full search)을 통해 최적 위치를 결정한다.

제한 조건-3 은 필터링 과정 및 최적 코드 탐색을 위한 오차 연산을 쉬프트(shift)와 덧셈 및 뺄셈으로만 표현할 수 있도록 하기 위해 정해진 값이다.

전송률을 낮추고, 고속 알고리듬을 적용하기 위해서는 부프레임에 사용할 최대 펄스의 개수를 제한해야 한다. 일반 사람의 피치 간격을 고려할 때 부프레임의 길이가 61 샘플일 경우 피치 성분은 3개를 넘지 않으며, 두 개일 경우에도 만족할 만한 성능을 얻을 수 있음을 실험을 통해 관찰하였다[7].

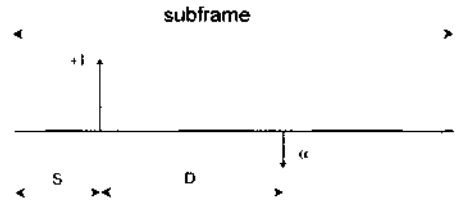

그림 4 두 개의 펄스로 구성된 구조화된 다중 펄스

그림 4에서 시작 펄스의 위치는 S, 두번째 펄스의 크기를 α, 그리고 펄스 사이의 간격을 D 라고 한다면, 구성된 코드는 (5) 식과 같이 표현할 수 있다.

$$c(n) = \delta(n-S) + \alpha\,\delta(n-S-D) \tag{5}$$

합성 필터의 충격 응답 함수를 $h(n)$ 이라고 하면, 필터링된 후의 신호 $q(n)$ 은 다음과 같이 표현된다

$$q(n) = \begin{cases} 0 & ,0 \le n < S \\ h(n-S) & ,S \le n < S+D \\ \alpha h(n-S-D) + h(n-S) & ,S+D \le n < N-1 \end{cases} \tag{6}$$

가중 필터를 통과한 신호를 $p(n)$ 이라고 한다면, 두 신호 사이의 오차 에너지 ϵ 은

$$\epsilon = \sum_{n=0}^{N-1}[\,p(n) - \beta q(n)\,]^2 \tag{7}$$

으로 주어지며, 오차를 최소로 하기 위한 이득 β 의 최적값

$$\beta_{opt} = \frac{\sum_{n=0}^{N-1}[\,p(n)q(n)\,]}{\sum_{n=0}^{N-1}[\,q^2(n)\,]} \tag{8}$$

으로 주어진다[3].

또한, β_{opt}를 (7) 식에 대입하면 최소 오차 에너지는 다음식과 같다

$$\epsilon_{min} = \sum_{n=0}^{N-1}[\,p^2(n)\,] - \frac{[\sum_{n=0}^{N-1}p(n)q(n)]^2}{\sum_{n=0}^{N-1}[\,q^2(n)\,]} \tag{9}$$

ϵ_{min} 의 앞 항은 상수값이므로 오차를 최소로 하기 위해서는 오른쪽 항을 최대로 하는 $q(n)$을 찾으면 된다.

필터링 과정은 (6)식과 같이 간단하게 표현될 수 있지만, 오차에너지를 최소로 하는 코드를 찾기 위해서는 부프레임 전체에 대한 세곱 계산과 상호 상관 연산이 필요하므로 연산량은 아직도 상당함을 알 수 있다. 이러한 문제는 고속 반복 알고리듬을 사용하여 해결할 수 있다[7].

핀스 개수 뿐만 아니라 두번째 펄스 탐색 영역 및 양자화 레벨도 성능 및 전송률 면에서 중요한 변수이다. 같은 전송률에서는 양자화 레벨보다는 탐색 영역이 더 중요한 변수이며, 제한된 크기에서는 양자화 레벨이 1 단계 이상 증가하더라도 별 차이가 없다[7]. 따라서, 두번째 펄스를 찾기 위한 탐색 영역 D 는 첫번째 펄스로부터 32 샘플 차이가 나는 지점까지로 제한하여 5 비트를 할당한다. 또한, 두번째 펄스의 크기는 2비트로 양자화 하여 4가지 값, {+1, -1, 0.5, -0.5} 을 갖도록 한다

나. 기타 구간

펄스의 모양이 일정하게 유지되는 유성음 구간에서는 위와 같은 구조화된 펄스를 이용하는 방법이 효율적이다. 그러나, 유성음에서 다른 형태의 유성음으로 변하는 부분이나 무성음 구간에서와 같이 랜덤 성분이 강하게 나타나는 부분에서는 2개의 펄스로는 모델링이 어려워지므로 성능이 저하된다. 여러 개의 펄스 조합으로 코드를 구성하면 이를 해결할 수 있지만, 전송률 및 계산량이 증가하게 되므로 벡터 양자화를 통한 코드북을 사용하는 것이 효율적이다.

다중 모드 부호화기의 평균 전송률을 낮추기 위해서는 무성음 구간에서는 장구간 예측을 수행하지 않는 것이 효율적이다.

그러나, 본 논문의 모드 2 에서와 같이 유/무성음 특징을 나타내는 구간을 모두 하나의 모드에 포함하는 경우에는 이 기능을 일관적으로 적용하기 어렵다. 추가 정보 없이 장구간 예측 여부를 결정하기 위해 본 논문에서는 구간마다 모드가 변화되는 상태를 살피는 방법을 이용하였다. 모드-B로 판별된 구간에서는 이전 구간의 모드를 검토하여 이전 구간이 모드-A일 경우에는 장구간 예측을 수행하고 모드-B일 경우에는 장구간 예측을 수행하지 않는 방법을 사용한다.

그림 5는 이러한 방법을 종합하여 모드 변화에 따른 장구간 예측 사용 여부 및 여기 신호 생성 방법을 상태도로 나타낸 그림이다.

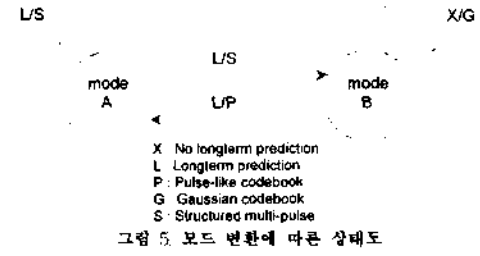

그림 5. 모드 변환에 따른 상태도

III. 실험 및 결과 고찰

제안한 부호화기의 성능 평가를 위해 모의 실험을 수행하였다. 실험에 사용된 데이타는 조용한 환경에서 남녀 화자 각 2명이 발음한 문장을 8KHz 샘플링하여 사용하였으며, 길이는 약 30 초 정도이다.

프레임의 길이는 32 msec로 구성하였으며, 부프레임은 프레임 길이의 1/4로 하였다. 프레임 길이를 2의 거듭 제곱 형태로 한 이유는 평형 상태의 유성음에 적용하는 펄스 위치 정보를 효율적으로 표시하기 위해서이다. 실험 결과 제안한 부호화기의 평균 전송률은 약 3.6 kbps로 나타났다.

성능 비교를 위해 4.8kbps 전송률을 갖는 DoD-CELP [14]를 사용하였으며, 평가 방법은 객관적 평가 방법으로 구간 신호대 잡음비(SegSNR)와 신호대 잡음비(SNR)를 사용하였으며, 주관적 평가 방법으로는 A-B Test 및 MOS (Mean Opinion Score)를 사용하였다

주관적 평가는 연구원 10명을 대상으로 실시하였으며, 평가 결과는 다음과 같다.

표 1. 객관적 성능 평가

	SegSNR	SNR
DoD - CELP	9.40	8.30
제안 방법	11.65	9.82

표 2. 주관적 성능 평가

DoD - CELP	제안 방법
46.0 %	54.0 %

(a) A-B Test

DoD-CELP	제안 방법
3.07	3.28

(b) MOS Test

제안된 방법은 평형 상태의 유성음 구간에 적용하는 구조화된 다중 펄스 방법과 같은 직접 탐색 방법의 장점으로 인하여 피치 성분을 정확히 모델링함에 따라 명료성이 뛰어났으며 이에 따라 주관적으로 좋은 평가를 받은 것으로 생각된다. 그러나, 무성음에서 유성음으로 변하는 부분과 유성음이 변화하는 부분에서는 다소 음질 저하 현상이 발생하였는데, 이는 펄스 모델이 단순화되고 코드북의 크기가 제한됨에 따라 발생하는 것으로서 좀 더 낮은 전송률에서는 고려해야할 문제라고 판단된다.

IV. 결 론

본 논문에서는 구간 음성의 특성에 따라 다른 여기 신호를 사용하여 4 kbps 이하에서 우수한 성능을 갖는 음성 부호화기를 제안하였다. 음성 신호의 단구간 에너지와 장구간 예측 후의 에너지를 이용하여 평형 상태 유성음(모드-A)과 기타 구간(모드-B)으로 나누고 모드 변화에 따라 각각 다른 여기 신호를 적용하였다. 모드-A 구간에는 구조화된 다중 펄스를 이용하였고, 모드-B 구간에는 이전 프레임의 모드에 따라 백색 가우시안 잡음 혹은 펄스형 잡음을 코드북으로 구성하여 사용하였다.

전체 성능 평가 결과 4.8 kbps DoD-CELP에 비해 구간 신호대 잡음비는 1~2 dB 정도 높은 수치를 나타내었으며, A-B Test와 MOS를 사용한 주관적 성능 평가도 우수한 성능을 나타내었다.

참 고 문 헌

[1] B. S. Atal, V. Cuperman and A. Gersho, *Advances in Speech Coding*, Kluwer Academic Publisher, 1991.

[2] *Speech and Audio Coding for Wireless and Network Applications*, Kluwer Academic Publisher, 1993.

[3] M. R. Schroeder, B. S. Atal, "Code Excited Linear Prediction(CELP) High-Quality Speech at Very Low Bit Rates," *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 25.1.1-25.1.4, 1985.

[4] K. Ozawa, M. serizawa, T. Miyano and T. Nomura, "M-LCELP Speech Coding at 4kbps," *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp.269-272, May 1994

[5] J. Martin, B. Wachter, "A CODEC Candidate for the GSM Half-Rate Speech Channel" *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1994.

[6] J. P. Campbell, T. A. Tremain, V. C. Welch, "DoD 4.8Kbps Standard (Proposed Federal Standard 1016)," *Advances in Speech Coding*, Kluwer Academic Publisher, 1991.

[7] 강 홍구, 서 정태, 차 일환, 윤 대희, "구조화된 다중 펄스 열을 이용한 낮은 전송률 음성 부호화기" 한국 통신학회 제출, 1995. 2.

[8] B. S. Atal, J. R. Remde, "A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates," *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 614-617, 1982.