저 전송률 음성 부호화기를 위한 여기 신호 개선 알고리즘에 관한 연구

이미숙*, 김홍국**, 최승호***, 김도영*

* 한국전자통신연구원 네트워크 연구소
** 광주과학기술원 정보통신공학과
*** 서울산업대 전자정보공학과

Enhancement of Excitation in Low-bit-rate Speech Coders

Mi Suk Lee*, Hong Kook Kim**, Seung Ho Choi***, and Do Young Kim*

* Network Technology Laboratory, Electronics and Telecommunications Research Institute
** Dept. of Information and Communications, Kwangju Institute of Science and Technology (K-JIST)
*** Dept. of Electronics and Information Eng., Seoul National University of Technology
E-mail : *[lms63425 dyk]@etri.re.kr, ** hongkook@kjist.ac.kr, ***shchoi@snut.ac.kr

Abstract

In this paper, we propose a new excitation enhancement technique to improve the speech quality of low bit rate speech coders. The proposed technique is based on a harmonic model and it is employed only in the decoding process of speech coders without any additional bits. We develop the procedure of harmonic model parameters estimation and harmonic generation, and apply the technique to a current state of the art low bit rate speech coder, ITU-T G.729 Annex D. Also its performance is measured by using the ITU-T P.862 PESQ score and compared to those of the phase dispersion filter and the long-term postfilter applied to the decoded excitation. It is shown that the proposed excitation enhancement technique can improve the quality of decoded speech and provide better quality for male speech than other techniques.

I. 서론

저 전송률 음성 부호화기에서는 음질을 높이기 위해 인간의 청각적 특성을 이용한 스펙트럼 포락선과 여기 신호 개선 알고리즘들이 개발되어 사용되고 있다 [1]. 스펙트럼 포락선을 개선시키는 방법 중 하나는 단구간 후처리 필터 (shot-term post-filter)로 스펙트럼의 피크 부분은 강조하고 널 부분은 더욱 깊게 하여 스펙트럼의 널 부분에서 발생하는 잡음에 의한 머플링 효과를 감쇄 시킨다 [2]. 또한 8kbps 미만의 전송률을 갖는 code-excited linear prediction (CELP) 유형의 음성 부호화기에서 적은 수의 펄스로 여기 신호를 모델링 할 때 발생하는 거친 소리는 phase dispersion 필터를 사용하여 완화시킬 수 있다 [1][3]. 그럼에도 불구하고 원래의 음성신호에 비해 하모닉 구조가 제대로 복원되지 않기 때문에 재생된 음성신호에 왜곡이 발생한다. 복호화된 음성신호의 하모닉 구조를 개선하기 위해서 하모닉 후처리 필터 또는 피치 후처리 필터를 사용하기도 한다 [2]. 그러나 이들 방법은 전 주파수 대역에 걸쳐 하모닉 구조를 변경하기 때문에 제대로 복원된 낮은 주파수 대역에 존재하는 하모닉 구조도 변경시킬 수 있다.

이 논문에서는 하모닉 모델 방법에 기반을 둔 여기 신호 개선 알고리즘을 제안한다. 이 알고리즘은 낮은 주파수 영역에 있는 하모닉 구조는 그대로 유지하면서 높은 대역에 존재하는 하모닉 구조를 개선시킨다. 이 알고리즘을 ITU-T 표준 G.729 Annex D [1]에 적용하여 기존의 여기 신호 개선 알고리즘과 성능을 비교하였다.

II. 기존의 여기 신호 개선 알고리즘

본 논문은 정보통신부가 출연한 선도기반 기술개발 연구사업에 의한 연구 결과물의 일부입니다.

저 전송률 음성 부호화기에서는 고정 코드북의 펄스 수가 작기 때문에 발생하는 문제점을 완화시키기 위해서 phase dispersion 알고리즘이나 피치 필터와 같은 여기 신호 개선 알고리즘을 사용할 수 있다. 그림 1은 일반적인 CELP 유형의 복호화기에 대한 블록도이다. 합성 필터의 입력신호는 적응 코드북과 고정 코드북에 의해 만들어진다. 복호화된 여기 신호 $e(n)$은

$$e(n) = g_p x(n) + g_c c(n) \tag{1}$$

으로 표현되며, $x(n)$과 $c(n)$은 각각 적응 코드북 (adaptive codebook)과 고정 코드북 (fixed codebook)이고 g_p와 g_c는 각 코드북의 이득 (gain)이다. 일반적으로는 $e(n)$을 합성필터의 입력으로 사용하지만, 저 전송률 음성부호화기에서는 $e(n)$을 개선 시킨 $e_h(n)$을 합성 필터의 입력으로 사용한다. 이 절에서는 $e(n)$을 개선하여 $e_h(n)$을 구하는데 사용되는 phase dispersion과 피치 필터에 대해 간단히 살펴보기로 한다.

2.1 Phase Dispersion

CELP 형태의 음성 부호화기에서 전송률을 줄일 수 있는 간단한 방법은 고정 코드북의 펄스 수를 줄이는 것이다. 그러나 고정 코드북의 펄스 수가 적어지면, 준 주기적인 성분이 복호화된 음성신호에 더해지는 것과 같은 영향을 주어 음질을 저하시키는 원인이 된다. 이러한 영향은 특히 3000 Hz 이상에서 두드러지며 유성음 보다는 무성음이나 배경잡음과 같이 신호에서 크게 나타난다. 고정 코드북의 펄스 수가 적기 때문에 나타나는 이러한 왜곡은 위상을 수정함으로써 제거할 수 있다고 한다. 위상을 dispersion 함으로써 이러한 왜곡을 완화시키는 알고리즘이 phase dispersion 이다. 다음 식은 $e(n)$으로부터 $e_h(n)$을 구하는 과정을 나타낸다.

$$e_h(n) = g_p x(n) + g_c c(n) * h_p(n) \tag{2}$$

여기서 $h_p(n)$은 phase dispersion filter 의 임펄스 응답이다. 식 (2)에서와 같이 $h_p(n)$은 고정 코드북에만 적용되며, 음의 특성에 따라 다른 값을 사용한다. 다시 말해, 유성음의 경우에는 위상을 너무 많이 수정하면 약간 잡음이 섞인 듯한 소리가 발생할 수 있다. 따라서 우선 현재 분석하고자 하는 프레임을 유성음과 무성음으로 분류한 후, 그 결과에 따라 phase dispersion의 정도를 다르게 한다. CELP 부호화기에서는 전송된 적응 코드북 이득 값에 따라 음성신호를 세 가지 유형으로 분류하여 위상을 dispersion 할 하위 주파수와 dispersion 정도를 결정한다. 예를 들면, 피치 이득이 0.9 이상이면 phase dispersion 을 하지 않고, 피치 이득이 0.5

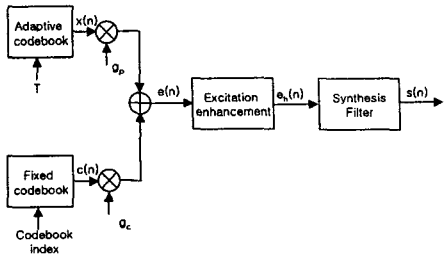

그림 1. 여기 신호 개선 알고리즘을 갖는 CELP 복호화기.

이하이면 2000 Hz 부터 π 만큼, 0.5에서 0.9 사이에서는 3000 Hz 부터 $\pi/2$ 만큼 위상을 dispersion 한다 [1][3].

2.2 Pitch Filtering

피치 필터는 피치 고조파들 사이에 있는 주파수 성분을 감쇄하여 피치의 주기성을 강조하기 위해 사용되는 필터이다. ITU-T 표준 G.729 Annex D 에서는 합성된 음성신호의 음질을 향상시키기 위해 식 (3)과 같은 형식의 피치 필터를 장구간 후처리 필터 (long-term postfilter)와 함께 사용한다. 하지만, 식 (3)의 필터는 기본적으로 comb filter 의 확장된 형태에 해당하며, 여기 신호의 주기성 향상에도 사용할 수 있다 [4].

$$H_p(z) = \frac{(1 + \gamma_p g_l z^{-T})}{1 + \gamma_p g_l} \tag{3}$$

여기서 g_l은 적응 코드북 이득에 비례 하는 값이고 γ_p는 피치 필터링의 정도를 나타내는 값으로 G.729 에서는 0.5 를 사용하고 있다. T 는 전송된 피치를 기반으로 새롭게 구한 피치 주기로 두 단계로 구해진다. 먼저, 첫 번째 프레임의 피치 값으로 부터 ±1 에 해당하는 구간에서 복호화된 여기 신호로부터 정수형 피치를 구한다. 그리고 이렇게 구해진 피치 근처에서 1/8 의 분해능으로 fractional 피치를 구한다.

III. 하모닉 모델 기반 여기신호 개선 알고리즘

하모닉 모델 기반 개선 기술은 그림 2 와 같은 과정에 의해 여기 신호의 고주파 영역 성분을 개선하는 것이다. 복호화된 여기 신호 $r(n)$과 피치 주기 T_0로 부터 개선된 여기 신호 $r_p(n)$을 발생시키며, 주파수 영역에서 개선된 여기 신호의 스펙트럼 $R_p(\omega)$는 식 (4)와 같이 표현된다.

$$R_p(\omega) \approx \begin{cases} R(\omega), & \omega < \omega_l \text{ or } \omega > \omega_h, \\ (1 - G_p(\omega))R(\omega) + G_p(\omega)R_h(\omega), & \omega_l \le \omega \le \omega_h \end{cases} \tag{4}$$

여기서 $G_p(\omega)$는 $r(n)$과 $r_p(n)$의 이득을 정규화하기 위한 것이다. 식 (4)는 시간영역에서 부 프레임 기반으로 실현된다. 우선, $0 \le n < N_s$ 구간에서 $r_h(n)$은 다음과 같이 표현된다.

$$r_h(n) = \sum_{m=M_l(T_0)}^{M_h(T_0)} A_m c_m(n; T_0, n_m), \tag{5}$$

$$c_m(n; T_0, n_m) = \cos((n+n_m)\frac{2\pi m}{T_0}), \tag{6}$$

여기서 N_s는 부 프레임의 크기이고 T_0는 적응 코드북의 피치주기이며, n_m과 A_m은 각각 현재 부 프레임에 대한 m 번째 사인파의 추정된 첫번째 위상과 크기이다. 그리고, $M_l(T_0)$와 $M_h(T_0)$는 주어진 T_0 에서 ω_l 과 ω_h 에 해당하는 하모닉의 인덱스이다. 결국 시간영역에서의 하모닉 개선은 다음과 같이 수행된다.

$$r_p(n) = r(n) + g_p r_h(n), \tag{7}$$

여기서 g_p 는 $G_p(\omega)$에 해당하는 하모닉 개선의 시간영역 이득이다. 식 (5)와 (6)에서의 위상과 크기를 추정하는 방법 및 $M_l(T_0)$와 $M_h(T_0)$을 설정하는 방법은 다음 세부 절에서 설명하도록 한다.

3.1 위상 추정

m 번째 사인파의 위상 파라미터인 n_m 은 입력 여기 신호와 식 (6)에서 정의된 사인파들 간의 상관도를 기반으로 한다. 즉, 주어진 m 에 대해서 n_m 은 $[0, \lfloor T_0/m \rfloor]$영역에 포함되며, 추정된 시간지연 n_m^* 은 다음의 수식에 의해서 얻어진다. 여기서 $\lfloor x \rfloor$는 x 보다 작거나 같은 정수를 의미한다.

$$n_m^* = \underset{0 \le k \le \lfloor T_0/m \rfloor}{\operatorname{argmax}} \sum_{n=0}^{N_T-1} r(n) c_m(n; T_0, k), M_l(T_0) \le m \le M_h(T_0) \tag{8}$$

식 (8)은 모든 m 에 대해서 n_m^* 을 추정하기 위해서 모든 부 프레임마다 $(M_h(T_0)\text{-}M_l(T_0)\text{+}1)$ 회 반복된다.

3.2 크기 추정

각 하모닉 주파수의 크기는 $r(n)$과 $r_h(n)$ 사이의 최소평균자승오차 (minimum mean-square error)를 기반으로 수행되며, 이렇게 함으로써 스펙트럼이 피크를 가지는 주파수를 보전 할 수 있다. A_m 은 다음 식의 오차를 최소화함으로써 얻어진다.

$$E = \sum_{n=0}^{N_T-1} (r(n) - r_h(n))^2 = \sum_{n=0}^{N_T-1} (r(n) - \sum_{m=M_l(T_0)}^{M_h(T_0)} A_m c_m(n; T_0, n_m))^2 \tag{9}$$

위 수식을 미분 함으로써 A_m의 최적 값인 A_m^*은 다음과 같다.

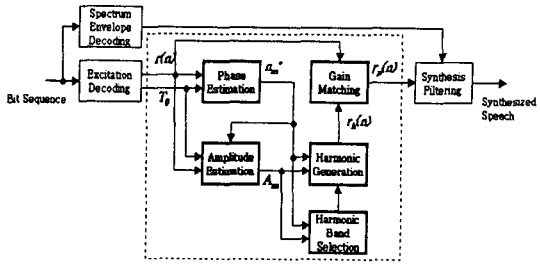

그림 2. 하모닉 기반 여기 신호 개선 과정.

$$A_m^* = \frac{\sum_{n=0}^{N_T-1} r(n) c_m(n; T_0, n_m^*)}{\sum_{n=0}^{N_T-1} c_m^2(n; T_0, n_m^*)} \tag{10}$$

여기서 $N_T(\le N_s)$ 는 T_0의 정수배가 N_s에 근접한 가장 큰 정수이다. 식 (10)에서 사인파의 직교성 (orthogonality)을 다음과 식 (11)과 같이 가정하며, 여기서 $M(T_0) \approx \lfloor T_0/2 \rfloor$이다

$$\sum_{n=0}^{N_T-1} c_m(n, T_0, n_m^*) c_k(n, T_0, n_k^*) = \delta(m-k), 1 \le m, k \le M(T_0) \tag{11}$$

실제로, $M_h(T_0)$ 부근에서 스펙트럼이 지나치게 개선되는 것은 $A_m^* = (1 - C_p \frac{m - (M_l(T_0)-1)}{M(T_0)}) A_m^*, \; M_l(T_0) \le m \le M_h(T_0)$ 을 적용하여 방지한다. 여기서, $M(T_0) = M_h(T_0) - M_l(T_0) + 1$ 이고 C_p는 고주파수 크기를 감축하기 위한 파라미터이며, 실험적으로 0.5로 고정한다.

3.3 하모닉 밴드 선택

본 논문에서는 하모닉의 최고 주파수 ω_h 은 $\lfloor T_0/2 \rfloor - 1$ 로 결정하였다. 최저 주파수 ω_l 은 매 부 프레임마다 다음과 같이 적응적으로 결정하였다. 3.1 절에서 위상 추정과 3.2 절에서의 크기 추정 결과를 이용하며, m 번째 하모닉 성분을 $h_m(n)$이라고 할 때, 이것과 원 여기 신호 $r(n)$과의 정규화된 상관도 ρ_m 을 $\rho_m = \sum_{n=0}^{N_T-1} r(n) h_m(n) \Big/ \sum_{n=0}^{N_T-1} r^2(n) \sum_{n=0}^{N_T-1} h_m^2(n)$ 과 같이 구한다. 이로부터 $d = \sum_{m=1}^{m^*} (1 - \rho_m)$가 임계치를 넘기 시작하는 m^*을 최저 주파수의 하모닉 인덱스로 결정한다.

3.4 하모닉 생성

현재 부 프레임에서 각 하모닉 주파수에 대한 추정된 위상과 크기는 이전 부 프레임과 내삽 (interpolation)되며, 이는 부 프레임 경계에서 스펙트럼의 변화를 부드럽게 하는 역할을 한다. 실질적으로 발생된 여기 신호 $r_h(n)$은 다음과 같이 표현된다.

$$r_h(n) = \sum_{m=M_1(T_0)}^{M_2(T_0)} A_m(n)\cos(\theta_m(n)) \quad (12)$$

여기서 [5]에 있는 스펙트럼 크기 및 위상의 내삽 과정을 사용하여 크기 $A_m(n)$과 위상 $\theta_m(n)$을 내삽하였다. 결과적으로 여기 신호의 개선된 부분 $r_h(n)$은 식 (12)를 사용하여 얻어진다. 마지막으로 이득 정규화 과정은 $\sum_{n=0}^{N_s-1} r_p^2(n) = \sum_{n=0}^{N_s-1} (r^2(n) + g_p r_h(n))^2 = \sum_{n=0}^{N_s-1} r^2(n)$ 의 관계로부터 $g_p = -2\sum_{n=0}^{N_s-1} r(n)r_h(n) / \sum_{n=0}^{N_s-1} r_h^2(n)$ 을 이용하여 얻어진다. 마지막으로 그림 2 에서 합성필터의 입력으로 $r_p(n)$이 사용된다.

IV. 성능평가

본 논문에서 제안된 하모닉 모델 기반 여기 신호 개선 알고리즘을 ITU-T 표준 G.729 Annex D에 적용한 후에 기존의 여기 신호 개선 알고리즘과 성능을 비교하였다. 성능 평가를 위해서 NTT-AT 데이터 베이스 (참고 문헌 [6])에 있는 한국어 음성 중에서 남, 여 각각 4 명의 화자가 발성한 약 6-8 초 정도 되는 음성 샘플을 이용하였다. NTT-AT 데이터베이스에 있는 샘플은 16 kHz로 샘플링 된 신호이며, ITU-T 표준 G.191 소프트웨어 툴 [7]을 이용하여 8 kHz 로 다운 샘플링 하였다. 그리고 성능 측정을 위해 perceptual evaluation of speech quality (PESQ) [8]를 이용하였다. PESQ를 구하기 위해 G.729에서 사용하고 있는 여기 신호 개선 알고리즘인 phase dispersion 을 pitch filter 와 본 논문에서 제안하고 있는 harmonic 개선 알고리즘으로 대체하였다.

한편, phase dispersion 은 주로 unvoiced harmonic 을 개선하는 특징을 가지고 있고, 반면, pitch filter 와 harmonic 개선 알고리즘은 유성음의 하모닉을 개선하는 특징이 있기 때문에, 이들을 결합하는 방식을 고려할 수 있다. 즉, Phase Dispersion+Pitch Filter 방식은 적응 코드북의 이득이 큰 경우에는 pitch filter 를 적용하고 그 외의 경우에는 phase dispersion 을 적용한 경우를 의미한다. 마찬가지로, phase dispersion 과 harmonic 개선 알고리즘도 결합된다. 이를 Harmonic+Phase Dispersion 이라 한다. 표 1 은 남 · 여 화자에 대해 각 방식을 적용한 후 얻은 PESQ 점수를 보여 준다. PESQ 점수는 직접적인 청취 실험 대신에 이에 상응하는 mean opinion score (MOS)를 나타내는 것으로 높은 값은 좋은 음질을 의미한다. 표 1에서 보면 여성 화자의 경우에는 Pitch Filter 나 Harmonic filter 를 사용하는 경우에 좀 더 나은 성능을 보임을 알 수 있다. 여성화자의 음성신호에 대해서

표 1. 각 방식에 대한 PESQ 점수 비교

여기신호 개선 방식	화자	
	여성	남성
Phase Dispersion	3.371	3.660
Pitch Filter	3.414	3.656
Pitch Filter + Phase Dispersion	3.409	3.651
Harmonic	3.378	3.677
Harmonic + Phase Dispersion	3.376	3.672

는 pitch filter가 남성화자의 음성신호에 대해서는 본 논문에서 제안하고 있는 하모닉 모델 기반의 여기 신호 개선 알고리즘이 좋은 성능을 보이고 있음을 알 수 있다.

V. 결론

본 논문에서는 저 전송률 음성 부호화기에서 나타나는 음질 저하를 줄이기 위해 하모닉 모델에 기반을 둔 여기 신호 개선 알고리즘을 제안하였다. 제안한 알고리즘은 여성화자 보다는 남성화자에서 좋은 성능을 보였으며, 기존의 ITU-T 표준 G.729 Annex D 에서 사용하고 있는 phase dispersion 알고리즘에 비해 나은 성능을 보여주고 있다. 본 논문에서 제안하고 있는 알고리즘은 전송된 피치 주기를 이용하여 하모닉을 발생시키므로 피치 검출의 정확성에 영향을 받을 수 있다. 따라서 피치 필터처럼 복호화 단에서 좀 더 정확한 피치를 사용한다면 성능 향상을 가져 올 수 있다.

참고문헌

[1] ITU-T Recommendation G.729 Annex D, *6.4 kbit/s CS-ACELP speech coding algorithm*, Sept. 1998.

[2] J. H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech Audio Process.*, vol.3, no. 1, pp. 59-71, Jan. 1995.

[3] R. Hagen, E. Ekudden, B. Johansson, and W. B. Kleijin, "Removal of Sparse-Excitation Artifacts in CELP," in *Proc. ICASSP*, Seattle, WA, pp. 145-148, May 1998.

[4] J. Lim, A. V. Oppenheim, and L. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoustic. Speech Signal Process.*, vol. 26, no. 4, pp. 354-358, Aug. 1978.

[5] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," IEEE *Trans. Acoustic. Speech Signal Process.*, vol. 36, no. 8, pp. 1223-1235, Aug. 1988.

[6] NTT-AT, *Multi-lingual speech database for telephonometry*, 1994.

[7] ITU-T Recommendation G.191, *Software tools for speech and audio coding standardization*, Nov. 2000.

[8] ITU-T Recommendation P.862, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, Feb. 2001.