4800 bps CELP 음성 부호화기에 적용한 대역폭 확장에 관한 연구

박 진 수, 김 형 순
부산대학교 전자공학과

A Study on the Bandwidth Extension Adopted for 4800 bps CELP Speech Coder

Jin Soo Park, Hyung Soon Kim
Dept. of Electronics Eng., Pusan National University
E-mail : {jinsp, kimhs}@pusan.ac.kr

Abstract

Most existing telephone networks transmit narrowband speech witch has been bandlimited below 4 kHz. Compared with wideband speech up to 8 kHz, narrowband speech shows reduced intelligibility and a muffled quality. Bandwidth extension is a technique to generate wideband speech by reconstructing 4-8 kHz highband speech without any additional information. This paper presents experimental results of the bandwidth extension adopted for 4800 bps CELP speech coder. In this experiment, we examine various methods for reconstruction of wideband spectrum and excitation signal, compare and analyze their performance by performing the subjective preference test and measuring the cepstral distortion.

I. 서 론

아날로그 전화망과 이동 통신망을 포함해 현존하는 대부분의 음성 통신 시스템은 대역폭이 300 Hz에서 3.4 kHz 영역으로 제한된 협대역 음성 신호의 전송을 기반으로 한다. 협대역 음성의 음질은 20 Hz에서 8 kHz의 대역폭을 갖는 광대역 음성의 경우와 비교했을 때, 명료성이 감소하고 억눌린 음질을 갖는다.

최근 들어 멀티미디어 환경이 급속히 발전하면서 사람들에게 고음질의 음성 및 음향은 매우 친숙하게 되었다. 이런 현상으로 기존의 통신 서비스의 음질을 향상시키기 위한 요구가 증대하고 있다. 전화 서비스의 관점에서 광대역 음성을 제공하는 것은 이러한 요구를 충족하는 한 방법이라고 할 수 있다.

대역폭 확장은 협대역 음성으로부터 저주파 성분(20 Hz - 300 Hz)과 고주파 성분(3.4 kHz - 8 kHz)을 복원하여 광대역 음성을 생성하는 기술이다[1][4][7]. 광대역 음성의 복원은 두 가지 가정을 전제로 구현된다. 첫째, 협대역 음성은 저주파 및 고주파 성분과 밀접한 연관 관계가 있다. 둘째, 저주파 및 고주파 성분의 복원이 완전하게 정확하지 않아도 음질을 상당히 높일 수 있다. 대역폭 확장의 장점은 전송된 음성 이외의 추가적인 정보가 없이도 광대역 음성의 복원을 통해 음질을 높일 수 있다는 점이다[1].

CELP(Code Excited Linear Prediction) 부호화기는 통신 산업에서 광범위하게 사용되는 대표적인 음성 부호화기 중의 하나이다[6]. 본 논문에서는 CELP 부호화기 가운데 최초로 표준화된 4.8 kbps FS-1016 부호화기를 기반으로 대역폭 확장에 관해 실험한 내용과 결과를 기술하였다. 본 연구에서는 CELP 부호화기의 디코더 내부에 대역폭 확장 모듈을 추가하여 대역폭을 6.8 kHz까지 확장시키는 것을 기본 실험 내용으로 하였다. 실험 결과로 여기 신호 생성 기법에 따른 선호도 조사 결과와 스펙트럼 복원 기법에 따른 스펙트럼 왜곡의 계산 결과를 나타내었다. 본 논문은 다음과 같이 구성된다. 2장에서 대역폭 확장에 대한 내용을 소개하고, 3장과 4장에서 실험 환경 및 실험 결과를 기

술한 후, 5장에서 결론을 맺는다.

II. 대역폭 확장

2.1 합성 시스템 구성

CELP 부호화기에 적용된 대역폭 확장은 그림과 같이 구성된다.

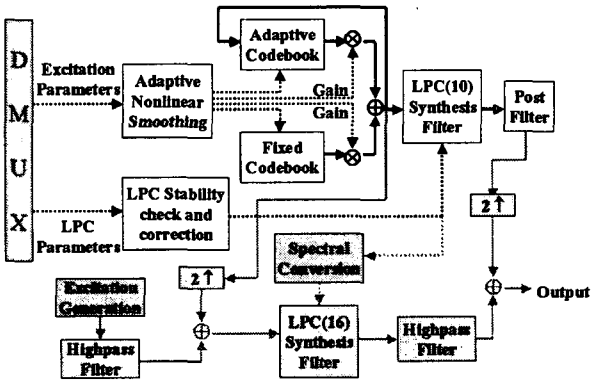

그림 1. 대역폭 확장을 적용한 CELP 디코더

그림은 FS-1016 CELP 부호화기의 디코더 내부에 대역폭 확장 과정이 결합된 것이다. 대역폭 확장 과정은 CELP 부호화기에서 사용되는 파라메타 중 Line Spectrum Pair(LSP)와 gain 등의 파라메타와 여기 신호를 이용하며 CELP의 프레임 처리 과정과 동기화되어 있다. 대역폭 확장은 기본적으로 광대역 스펙트럼 포락선을 복원하는 부분과 여기 신호를 생성하는 부분으로 요약된다. 스펙트럼 포락선의 복원은 CELP의 합성 필터를 구성하는 10차 LSP로부터 코드북 매핑, Gaussian Mixture Model(GMM) 기반 스펙트럼 변환 등의 기법을 통해 광대역 음성의 합성 필터를 구성하는 16차 LSP로 변환하는 과정이다. 광대역 합성 필터에 입력되는 여기 신호는 스펙트럼 folding, 코드북 매핑 등의 기법을 통해 생성된 고주파 성분과 CELP에서 생성되는 협대역 여기 신호를 더한 후 두 신호의 gain 보상 과정을 거쳐 생성된다. 광대역 합성 필터의 출력은 고역 필터를 거쳐 CELP의 협대역 출력 음성과 더해져 최종적으로 광대역 음성을 생성하게 된다.

2.2 스펙트럼 복원 기법

광대역 음성의 스펙트럼을 복원하기 위한 방법으로는 코드북 매핑, 통계적 복원, 선형 매핑 등을 포함한 여러 기법들이 있다. 본 논문에서는 벡터 양자화(VQ)를 이용한 코드북 매핑[1]과 통계적 복원 기법의 하나인 GMM 기반 스펙트럼 변환[2][3][7]을 적용하였다. 코드북 매핑은 협대역 및 광대역 스펙트럼 벡터로부터 각각의 코드벡터간에 일대일 대응을 이루도록 훈련된 한 쌍의 코드북을 사용한다. 합성 과정에서 협대역 스펙트럼 벡터가 입력되면 협대역 코드북으로 양자화된 후 코드인덱스가 결정된다. 동일한 코드인덱스를 사용해 광대역 코드북으로부터 결정된 코드벡터를 광대역 스펙트럼으로 사용하게 된다. 협대역 및 광대역 코드북을 훈련하기 위해서는 먼저 광대역 음성 데이터와 저역 필터를 통해 협대역 음성 데이터를 생성한다. 그리고 각각의 스펙트럼 포락선을 표현하는 LSP 또는 LPC 켑스트럼 등의 특징 인자를 추출한다. 광대역 특징 인자로부터 벡터 양자화 훈련을 통해 광대역 코드북을 생성하고 양자화를 통해 광대역 데이터의 코드북 인덱스를 얻는다. 동일한 시간 정보를 갖는 협대역 데이터를 광대역 데이터의 코드북 인덱스에 따라 클러스터링을 해서 협대역 코드북을 생성한다.

GMM 기반 스펙트럼 변환 기법은 스펙트럼 벡터의 확률 모델로부터 스펙트럼 변환 함수를 유도하여 대역폭 확장에 적용한 방법이다. 협대역 음성의 스펙트럼 벡터

$$x = [\, x_1\ x_2 \ldots x_m \,]^T \tag{1}$$

및 광대역 음성의 스펙트럼 벡터

$$y = [\, y_1\ y_2 \ldots y_n \,]^T \tag{2}$$

의 조합

$$z = [\, x_1\ x_2 \ldots x_m\ y_1\ y_2 \ldots y_n \,]^T \tag{3}$$

의 joint density는 Q개의 p(=m+n)-variate Gaussian 함수로서 다음 식과 같이 표현된다.

$$p(z \mid \alpha_i, \mu_i, C_i) = \sum_{i=1}^{Q} \frac{\alpha_i \exp[\, g_i(z)]}{2\pi^{p/2} \mid C_i \mid^{1/2}} \tag{4}$$

여기서

$$g_i(z) = -\frac{1}{2}\,(z - \mu_i)^T C_i^{-1} (z - \mu_i)$$

이다. 식에서 α_i, μ_i, C_i 는 각각 i 클래스의 사전 확률, 평균벡터 및 공분산행렬을 나타낸다. 협대역 스펙트럼 벡터와 광대역 스펙트럼 벡터 사이의 평균자승 오차를 최소화하는 매핑 함수는 다음 식과 같다.

$$F(x) = E[y \mid x]$$

$$= \sum_{i=1}^{Q} h_i(x)[\, \mu_i^{y} + C_i^{yx} C_i^{xx-1} (x - \mu_i^{y})] \tag{5}$$

여기서

$$h_i(x) = \frac{\dfrac{\alpha_i}{(2\pi)^m \mid C_i^{xx} \mid^{1/2}} \exp[\, g_i(x)]}{\displaystyle\sum_{j=1}^{Q} \frac{\alpha_j}{(2\pi)^m \mid C_j^{xx} \mid^{1/2}} \exp[\, g_j(x)]}$$

$$C_i=\begin{bmatrix} C_i^{xx} & C_i^{xy} \\ C_i^{yx} & C_i^{yy} \end{bmatrix},\ \mu_i=\begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}$$

이다.

GMM의 파라메타를 추정하는데 널리 사용되는 방법은 Maximum Likelihood(ML) 추정이며, 본 실험에서는 잘 알려진 Expectation and Maximization(EM) 알고리즘을 사용하여 파라메타를 추정하였다[2].

2.3 여기 신호의 생성 기법

복원된 스펙트럼 포락선과 함께 광대역 음성을 합성하는데 사용될 여기 신호의 생성 방법으로는 비선형 변환, 코드북 매핑, 스펙트럼 folding, 정현파 모델 등을 포함해 다양한 기법들이 있다. 본 논문에서는 3.4 kHz 이상의 고주파 성분의 여기 신호를 잡음으로 모델링한 방법과 스펙트럼 folding 및 코드북 매핑 기법을 적용하였다. 여기 신호가 잡음으로 모델링된 경우는 CELP의 코드북 인덱스로 광대역 잡음 코드북을 검색한 후 협대역 여기 신호에 더해지게 된다. 스펙트럼 folding은 8 kHz 샘플링된 협대역 여기 신호를 16 kHz로 upsampling하는 과정에서 일반적으로 행해지는 저역 필터링을 생략함으로써 구현된다. 여기 신호의 코드북 매핑은 스펙트럼 코드북 및 GMM의 훈련 과정에서 각각의 코드벡터 및 mixture에 대응하는 원음성의 여기 신호를 대표값으로 하여 코드북을 생성하여 합성에 사용하는 방법이다[1]. 코드북 매핑은 원음성의 여기 신호를 왜곡 없이 사용하기 때문에 스펙트럼 코드북이나 GMM의 훈련이 정교할수록 다른 방법에 비해 좋은 결과를 낼 것으로 예상할 수 있다.

III. 실험 환경

실험을 위해 국어공학연구소에서 구축한 PBS 589 문장에 대한 남녀 50명분의 발성 데이터 약 13시간 분량을 훈련 및 테스트에 사용하였다. VQ 코드북 생성 및 GMM 파라메타 추정 등 훈련용으로 40명분의 데이터베이스를 사용하였고 합성 테스트 및 성능 평가용으로 10명분의 데이터베이스 중에서 1명당 10문장씩 총 100문장을 사용하였다. 광대역 음성 신호는 저역 필터링을 통해 0 - 6.8 kHz의 대역폭을 갖도록 하였고, 협대역 음성 신호는 저역 필터링을 통해 0 - 3.4 kHz의 대역폭으로 제한하여 생성하였다.

협대역 및 광대역 음성의 스펙트럼 벡터로서 각각에 대해 10차 및 16차 LPC 분석을 한 후, 추출된 LPC 계수로부터 변환된 LSP 및 LPC 켑스트럼을 사용하였다. VQ 코드북의 크기와 GMM의 mixture 수는 모두 128로 하였다.

IV. 실험 결과

4.1 성능 평가 방법

대역폭 확장의 성능 평가를 위한 방법으로 원래 음성 신호와 합성 음성 신호 사이의 스펙트럼 왜곡을 측정하였다. 스펙트럼 왜곡의 표현식은 다음과 같다.

$$SD=\sqrt{\frac{20^2}{2\pi}\int_{-\pi}^{\pi}\left[\log\frac{|H(e^{j\omega})|}{|H^r(e^{j\omega})|}\right]^2 d\omega} \quad (6)$$

여기서 $H(z)$와 $H^r(z)$는 각각 원음성과 복원된 음성에 대한 LPC 필터의 전달함수이다. cepstrum을 이용한 스펙트럼 왜곡의 계산식은 다음과 같이 유도된다.

$$SD=\sqrt{2\cdot 10^2(\log e)^2\sum_{n=1}^{\infty}[c_n-c^r{}_n]^2} \quad (7)$$

본 실험에는 40차의 켑스트럼 계수를 스펙트럼 왜곡을 계산하는데 사용하였다.

4.2 성능 평가 결과

광대역의 원음성으로부터 추출한 스펙트럼과 원래의 여기 신호 및 각각의 생성 기법에 의한 여기 신호를 이용한 실험을 통해 여기 신호 생성 기법에 따른 선호도를 조사하였다.

표 1. 여기 신호에 따른 합성음의 선호도 평가 결과
(A=original, B=noise, C=folding, D=mapping)

		비교 대상				
		A	B	C	D	평균
선호 기준	A	-	10	7	9	**8.7**
	B	0	-	3	2	**1.7**
	C	2	5	-	3	**3.3**
	D	0	2	4	-	**2.0**

표 1은 원음성의 여기 신호(A)를 포함해 각각의 기법들(B, C, D)에 의해 합성된 음성 신호에 대한 선호도를 표시한 것이다. 선호도 조사는 예를 들어 A, B, C, D 중 A와 B를 들려 준 후 둘 중 선호하는 것에 1점을 주고 구분이 어려울 경우 0점을 주는 방식으로 총 6쌍의 평가음성에 대해 10명이 참가하였다. 예상

가능한 결과로서, 원음성의 여기 신호에 의한 합성음이 다른 방법들에 비해 월등한 높은 선호도를 나타내었다. 여기 신호 생성 기법에 따른 결과는 스펙트럼 folding에 의한 합성음이 비교적 좋은 것으로 나타났다.

위의 결과를 바탕으로 스펙트럼 folding에 의한 여기 신호 생성 기법을 사용하고 스펙트럼 복원 기법 및 특징 인자에 따른 실험을 하였다. 표 2는 원음성의 스펙트럼 및 LSP와 LPCC를 기반으로 VQ 코드북 매핑과 GMM에 의한 스펙트럼 복원 기법에 따라 스펙트럼을 복원했을 때의 결과를 비교한 것이다.

표 2. 스펙트럼 복원 방식에 따른 스펙트럼 왜곡

	all frames dB	<3dB dB (%)	>3dB dB(%)
original excitation	3.07	2.27 (50.7)	3.88 (49.3)
original	3.74	2.47 (29.0)	4.26 (71.0)
VQ(LSP)	5.31	2.65 (5.3)	5.45 (94.7)
GMM(LSP)	4.98	2.65 (7.3)	5.17 (92.6)
VQ(LPCC)	5.37	2.66 (4.7)	5.50 (95.3)
GMM(LPCC)	5.08	1.45 (11.5)	5.55 (88.5)

표의 제일 상단의 결과는 원음성의 스펙트럼과 함께 원음성의 여기 신호를 함께 사용한 경우의 결과로서 스펙트럼 복원 및 여기 신호 생성 기법에 따른 성능의 상한선으로 생각할 수 있다. 그 아래 결과는 원음성의 스펙트럼과 함께 스펙트럼 folding에 의해 생성된 여기 신호를 사용한 경우의 결과를 나타낸다. 스펙트럼 복원과 특징 인자의 관점에서는 LSP가 LPCC에 비해 조금 더 좋은 결과를 나타냈다. 코드북 매핑에 의한 스펙트럼 복원의 경우 원음성의 스펙트럼을 사용한 결과와 비교해 스펙트럼 왜곡이 대략 1.6 dB 더 높은 것은 코드북 훈련 과정에서 비롯된 VQ 왜곡의 결과라고 할 수 있다. 같은 맥락에서 GMM 기반의 스펙트럼 복원 방법은 동일한 훈련 데이터에 대해 코드북 매핑보다 더 좋은 성능을 내는 것을 확인할 수 있다. 이와 함께 훈련 데이터를 증가시켰을 때의 성능 변화를 확인하는 실험을 실시한 결과 훈련 데이터의 양이 성능에 미치는 영향을 확인할 수는 있으나 훈련 데이터의 증가율에 비해 스펙트럼 왜곡이 현저히 감소하지는 않았다. 이 것은 훈련 데이터의 양이 스펙트럼 왜곡에 미치는 영향의 주요인은 아니라는 의미로 해석된다. 그보다는 코드북의 크기가 스펙트럼 왜곡에 더 큰 영향을 미칠 것으로 생각된다. 향후 실험에서 코드북 크기가 스펙트럼 왜곡에 주는 영향을 확인할 계획이다.

V. 결 론

본 논문에서는 CELP 부호화기를 기반으로 대역폭 확장에 관한 실험 내용 및 결과를 기술하였다. 스펙트럼 복원 방법으로 VQ 기반 코드북 매핑과 GMM 기반 스펙트럼 변환을, 그리고 여기 신호 생성 방법으로는 잡음, spectral folding, 코드북 매핑 방법을 적용하였다. 실험 결과 스펙트럼 복원 방법에서는 코드북 매핑에 비해 GMM 기반 스펙트럼 변환 기법이 조금 더 우수한 결과를 나타내었다.

향후 실험에서는 GMM 기반 스펙트럼 변환 기법의 훈련을 강화하고, 코드북 및 mixture 크기 변화에 따른 스펙트럼 왜곡 변화를 확인하며, CELP 부호화기와 대역폭 확장의 더욱 효율적인 결합 방법을 모색해 볼 계획이다.

참고문헌

[1] Y. Yoshida and M. Abe, "An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping," Proc. of ICSLP94, pp. 1591-1594, 1994.

[2] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE trans. Speech and Audio Processing, Vol.3, no. 1, pp. 72-83, Jan. 1995.

[3] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," Proc. of IEEE ICASSP98, pp. 285-288, 1998.

[4] J. Epps and W. H. Holmes, "A new technique for wideband enhancement of coded narrowband speech," IEEE Workshop on Speech Coding, Porvoo, Finland, 1999.

[5] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," IEEE Trans. on Communications, Jan. 1980.

[6] M. R. Schoreder and B.S. Atal, "Code-excited linear prediction(CELP): High-quality speech at very low bit rates," ICASSP, pp. 937-940, 1985.

[7] K. Y. Park and H. S. Kim, "Narrowband to wideband comversion of speech using GMM based transformation," Proc. of ICASSP2000, Vol. 3, pp. 1843-1864, 2000.