음성압축을 위한 전처리기법의 비교 분석에 관한 연구

A Study on a Analysis and Comparison of Preprocessing Technique for the Speech Compression

장 경 아* · 민 소 연** · 배 명 진*
KyungA Jang · SoYeon Min · MyungJin Bae

ABSTRACT

speech coding techniques have been studied to reduce the complexity and bit rate but also to improve the sound quality. CELP type vocoder, has used as a one of standard, supports the great sound quality even low bit rate. In this paper, the preprocessing of input speech to reduce the bit rate is the different with the conventional vocoder. The different kinds of parameter are used for the preprocessing so this paper is compared with theses parameters for finding the more appropriate parameter for the vocoder. The parameters are used to synthesize the speech not to encode or decode for coding technique so we proposed the simple algorithm not to have the influence on the processing time or the computation time. The parameters in used the preprocessing step are speaking rate, duration and PSOLA technique.

Keyword : preprocessing technique, speaking rate, duration time, PSOLA

1. 서 론

G.723.1 보코더는 인터넷 폰이나 화상회의, voice mail system, voice-pager 등에 응용이 가능하며 현재 상용버전으로 나와 사용되고 있다[2]. 이 중 G.723.1은 5.3/6.3 kbps의 이중 전송률을 갖는 구조로 되어 있다[1]. 최적의 전송 환경을 위하여 두 개의 전송률을 사용하기 때문에 다른 보코더 표준안들에 비해서 더욱 응용성이 높다. 그러나 G.723.1 역시 음성신호를 성분 분리하여 합성하는 방식인 CELP 보코더 계열의 합성에 의한 분석방법을 사용하기 때문에 많은 계산량으로 인한 처리 시간의 소모를 피할 수 없다는 문제점을 갖고 있다. G.723.1은 두개의 서로 다른 보코더를 포함하고 있어 DSP칩으로 구현시 많은 내부 메모리와 계산량을 필요로 한다. 논문에서는 G.723.1 5.3 kbps ACELP를 기반으로 하여 음질을 유지하면서 전송률을 낮출 수 있는 새로운 부호화 방법을 소개하고, 이 방법들의 분석하고자 한다. 본 논문에서는 음성 데이터를 G.723.1 보코더 입력하기 전에 전처리단을 이용한다. 전처리단에 응용되는 기술은 기존의 파형 압축방법과는 전혀 다른 피치단위로 파형을 부호화하는 밥법, 지속시간과 발성률

* 숭실대학교 정보통신공학과
** 숭실대학교 전자공학과

을 고려한 방법, PSOLA 기법을 사용한 방법을 차례로 설명한다.

2. PSOLA 기법을 이용한 전처리 과정

2.1 NAMDF에 의한 포만트 유사도 측정

현재 프레임의 피치를 측정하는 방법으로는 다음과 같이 NAMDF를 정의하여 사용할 수 있다[3].

$$NAMDF(d) = -\frac{\sum_{n=1}^{N} |s(n) - s(n-d)|}{\sum_{n=1}^{N} |s(n)| + |s(n-d)|} \tag{1}$$

여기서 s(n)은 음성신호이고 N은 NAMDF를 구하려는 윈도우 구간이다. 지연인자 d를 점차 증가시키면서 NAMDF를 구해보면, 지연인자가 프레임 내 음성피치에 정수배가 될 때마다 NAMDF는 거의 영이 된다.

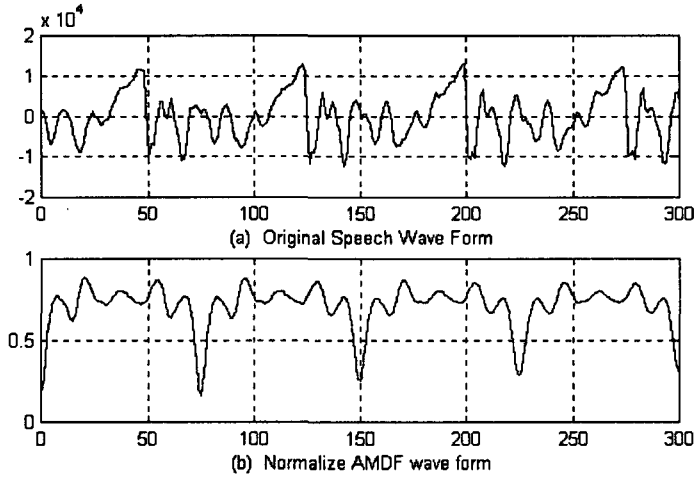

그림 1. (a) 음성파형 (b) NAMDF 파형

자기상관함수와 AMDF를 취했을 경우 영점 위치를 살펴보면 자기상관함수의 정확한 피크 값을 찾는 것이 AMDF의 피크 값을 찾는 것보다 더 어렵다는 것을 볼 수 있다. 이러한 이유 때문에 피치검색시에 잘못된 피크 값을 얻게 됨으로써 피치검색시 오차를 발생시킬 수 있는 문제를 내포하고 있어 AMDF가 자기상관함수 대신에 주기성을 강조하는데 오랫동안 적용되어 왔다[5]. 또한 AMDF는 곱셈을 사용하지 않는 장점이 있다. 단 규준화시 한 번의 나눗셈은 전체 계산량에 커다란 영향을 주지 않기 때문에 NAMDF의 장점을 유지할 수 있다. 본 논문에서는 NAMDF를 이용하여 피치를 검색하고 유사도 측정 구간을 정하였다. 그리고 한 구간 안

의 피크들의 변화는 Cross NAMDF법을 이용하여 측정할 수 있다. 본 논문에서는 Cross NAMDF법을 이용하여 포만트 유사도 측정에 적용하였다. 유성음 구간을 관찰하면 피치가 일정하게 유지되는 구간에서도 포만트는 조금씩 변화하는 것을 알 수 있다. 이러한 포만트의 정보는 한 피치주기 사이에 나타나는 피크의 수와 모양, 크기, 위치 등에 좌우된다. 따라서 포만트의 유사도를 측정하기 위하여 기준 피치와 인근 피치 주기 내에 나타나는 피크들의 특성을 비교하였다. 한 주기 안에 나타나는 피크들의 특성을 비교하기 위하여 기준피치와 인근피치 한 주기 파형에 대해 Cross NAMDF를 수행하였다.

$$NAMDF_{Cross}(d)=\frac{\sum_{n=1}^{N}|S_{ref}(n)-S_p(n-d)|}{\sum_{n=1}^{N}|S_{ref}(n)|+|S_p(n-d)|} \tag{2}$$

Cross NAMDF는 식 (2)과 같다. 여기서 S_{ref} 는 기준이 되는 피치주기의 파형이고 S_p는 p 번째 주기의 파형이다. N은 윈도우 크기이고 S_{ref}와 S_p 길이 중 작은값이다. d는 지연인자이다. 구해진 파형에 대한 면적은 식 (3)와 같이 구해진다.

$$A(p) = \frac{1}{N}\sum_{d=1}^{N} NAMDF_{Cross}(d) \tag{3}$$

여기서 $A(p)$는 p 번째 파형의 Cross NAMDF 파형의 면적이다. 구해진 면적과 기준 피치주기의 NAMDF 파형의 면적을 비교하여 유사도를 측정한다. 유사도 측정은 식 (4)과 같다.

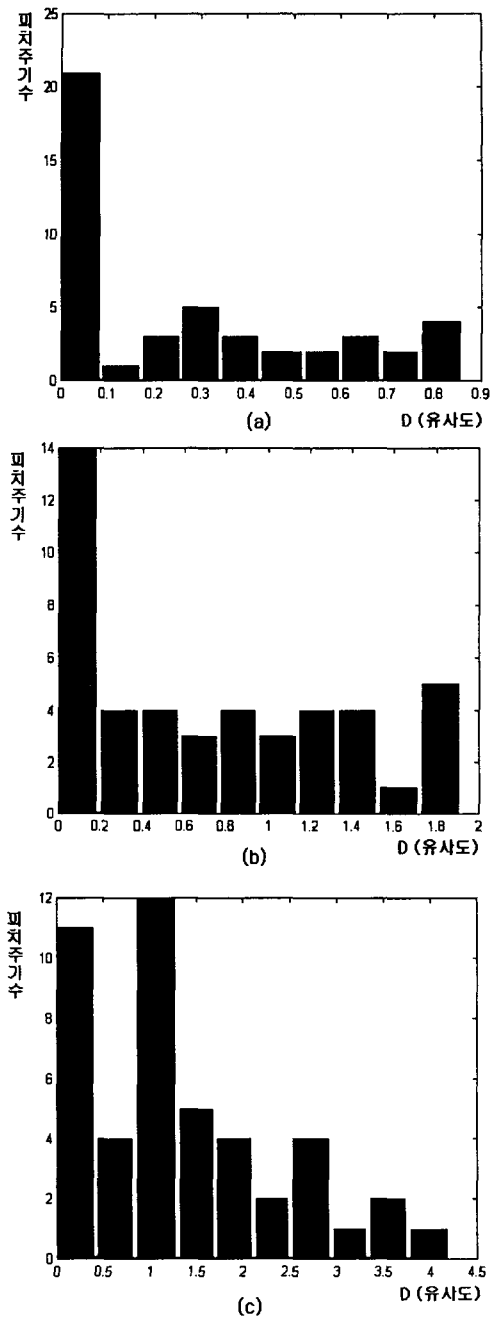

그림 2. 문턱값에 따른 피치주기의 압축률

첫 막대(D=0) → 전송되는 피치 주기 수, 그 외 (D>0) → 압축되는 피치 주기 수

(a) D = 1를 문턱값으로 했을 때 (45.6%)

(b) D = 2를 문턱값으로 했을 때 (30.4%)

(c) D = 5를 문턱값으로 했을 때 (23.9%)

$$D(p) = \frac{| A_{ref} - A_p |}{A_{ref}} \times 100 \quad (\%) \tag{4}$$

A_{ref}는 기준파형의 NAMDF 파형의 면적이고, A_p는 식 (3)와 같이 구한 기준파형과 인근 파형의 Cross NAMDF 파형의 면적이다. D(p)는 p 번째 파형의 포만트 유사도를 나타내며, 값이 작을수록 p 번째 파형은 기준 파형과 유사하다.

2.2 PSOLA 기법에 의한 음성합성

본 논문에서는 음성신호를 복원할 때 스펙트럼 왜곡률과 복잡성이 적은 PSOLA 방법이 적합하다[6]. 전송 또는 압축된 파형과 진폭정보와 피치정보를 이용하여 PSOLA 합성을 수행한다[7]. 그림 3은 PSOLA 기법으로 합성하는 과정을 나타내었다. G. 723.1 보코더 입력단에 들어가기 전에 NAMDF를 이용하여 포만트 유사도를 측정하여 기준파형과 유사도 적은 파형들의 차이값을 가지고 압축한다. 압축된 파형은 G.723.1 보코더에서 통과한 후 PSOLA 합성방식을 이용하여 음성파형을 복원한다.

그림 4은 G.723.1 보코더에 입력하기 전 전처리과정을 나타내는 블록도이다. 송신단에서는 먼저 한 프레임에 대한 NAMDF법을 사용하여 피치를 구한다. 피치는 그림 3의 (b)에서 가장 먼저 0점에 가까워지는 Valley까지의 간격으로 정한다. 이렇게 구해진 피치에 일치하는 한 주기를 기준 파형으로 정하고 저장하거나 전송한다. 기준 파형의 진폭정보를 추출하고 기준 파형만의 NAMDF를 수행하여 기준면적을 구한다. 기준면적은 유사도가 문턱값을 넘어 기준 파형이 달라질 때 갱신된다. 기준파형의 면적이 구해지면 처리된 파형의 피치만큼 전진하여 새로운 프레임을 잡고 NAMDF를 수행하여 피치를 구하고 진폭정보를 추출한다. 그림 5의 (a)는 원래 음성 파형이고, (b)는 피치정보를 표시한 그림이고, (c)는 합성을 위한 한 주기 파형이며. (d)는 PSOLA 방법을 이용하여 합성한 파형이다.

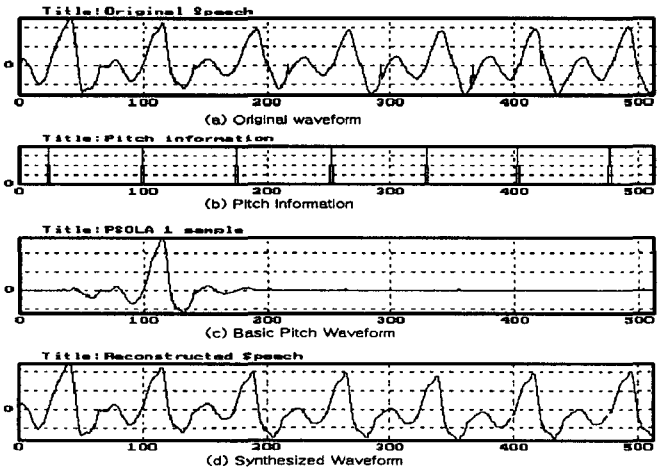

그림 3. 피치단위의 처리과정 예

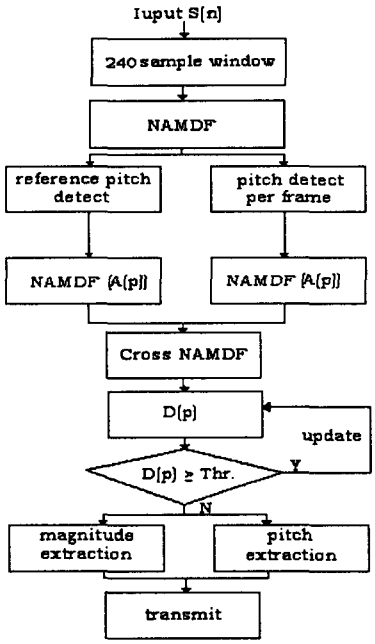

그림 4. 제안한 방법의 블록다이어그램

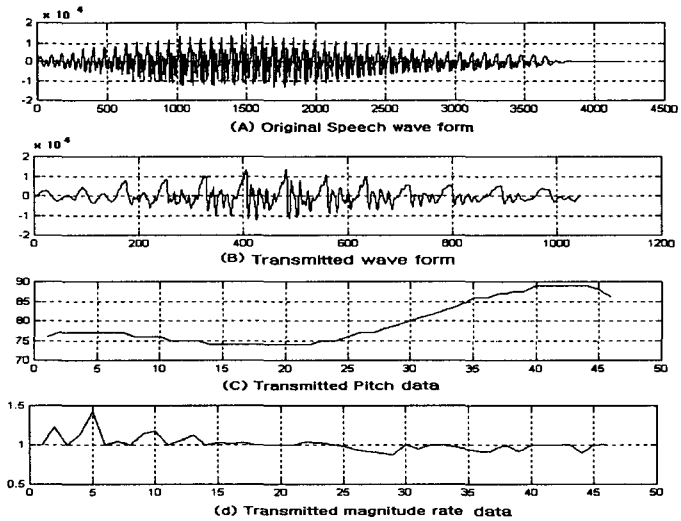

그림 5. '아' 음성에 대한 부호화

(A) 음성파형 (B) 전송되는 파형
(C) 전송되는 피치정보 (C) 전송되는 진폭(변화)정보

합성단에서는 전송된 파형과 피치정보, 진폭정보를 이용하여 PSOLA 방법으로 복원해낸다 송신단에서 문턱값을 변화시킴으로써 압축률을 조정할 수 있다.

3. 지속시간 변경을 적용한 전처리 과정

본 논문에서는 보코더 단에 입력 전에 지속시간을 변경하는 방법으로 FFT변환 특성을 이용해 음색의 변경 없이 지속시간을 변경하는 방법을 사용하였다. 본 방법은 주파수 영역에서의 지속시간 변경 법으로 FFT를 이용하여 계산시간을 줄이고 진폭과 위상에 각각 $1/2^n$ 배의 Decimation을 수행한 다음 G.723.1 보코더 입력하여 부호화시킨 후 G.723.1 복호화 단을 통과 후 FFT point의 $1/2^n$ point로 IFFT 과정을 수행함으로써 스펙트럼의 변경 없이 지속시간을 변경, 음성을 합성하였다. 우선 Frame 단위로 Segment된 음성신호를 FFT 통해 진폭성분과 위상성분으로 나눈다. 프레임 단위의 음성신호에 대하여 DFT를 하면 아래의 식과 같다.

$$S_{fr}(k) = \sum_{n=0}^{N-1} s(n) e^{-j2\frac{\pi}{N}kn}, \quad 0 \le n \le N-1 \tag{5}$$

이 신호는 실수부와 허수부로 나눌 수 있는데 아래의 식과 같다.

$$S_{fr}(K) = Re[S_{fr}(K)] + jIm[S_{fr(K)}] \tag{6}$$

진폭과 위상은 다음 식과 같이 정의된다.

$$M(K) = 10 \log S_{fr}^2(K) \tag{7}$$

$$\varphi(K) = \tan^{-1}(Im[S_{fr}(K)]/Re[S_{fr}(K)]) \tag{8}$$

FFT된 음성의 진폭과 위상스펙트럼에 원 신호의 1/2 배로 Decimation하기 전과 Decimation 한 후의 진폭과 위상 스펙트럼으로 나타낼 수 있다. 각각 진폭과 위상스펙트럼은 음질에 크게 영향을 주지 않는 범위 내에서 거의 변하지 않았으며 비율만 1/2 배로 된 것을 알 수 있다. 이렇게 얻어진 각 진폭과 위상선분에 $1/2^n$ 배로 주파수 축에서 Decimation 과정을 수행한다. 그런 다음 $1/2^n$ 배 point로 FFT를 수행하여 지속시간이 1/2정도 줄어든 음성을 얻어낼 수 있었다. 그림 3-1은 블록다이어그램을 나타낸다.

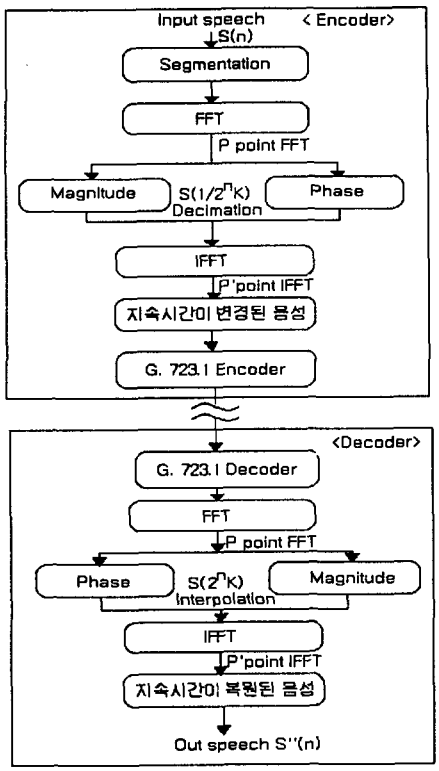

그림 6. 제안한 알고리즘의 블록도

4. 발성률을 적용한 전처리 과정

다음의 파라미터는 LSP 변화도를 이용한 발성율과 PSOLA 기법의 압축 파라미터이다. 여기서 발성률 측정하고자 하는 방법은 기존의 연구와 달리 처리시간과 알고리즘 자체가 간단하게 측정할 수 있도록 하였고 PSOLA 기법은 PSOLA 방법 중 TD-PSOLA 방법을 사용하였다. 보코더 자체 내부의 처리이나 계산량이 많으므로 인해 처리시간이 많이 소모되어, 적용한 알고리즘을 적용시킴으로써 보코더 내에서 계산량이나 처리시간을 더 부가시키지 않을 수 있는 알고리즘을 사용하였다. 본 논문에서 고려하는 발성속도는 묵음 부분이 제거된 음성신호에서의 발성속도이다. 본 논문에서는 먼저 묵음구간의 에너지와 LSP 파라미터를 정보를 이용하여 음성 검출을 수행하고, 파라미터를 추출하는 묵음구간은 발성시료의 처음부분을 이용하였다. 60 msec 동안의 평균 LSP 값을 사용하여 거리를 측정하기 위해 유클리디안 거리측정법을 사용하였다.

$$D(n) = \frac{1}{P}\sum_{i=0}^{P} |LSP_n(i) - LSP_{n+1}(i)|^2 \tag{9}$$

D(n)는 n 번째 분석구간과 n+1 번째 분석구간과의 LSP 거리를 나타내고, P는 LSP 분석 차수이다. 입력음성의 발성속도를 계산하기 위해서는 먼저 현재 처리되는 분석구간이 묵음인지 판정해야 된다. 묵음의 판정은 미리 구한 에너지 문턱값과 LSP 파라미터를 이용한다. 묵음 판정이 끝난 후 인접 분석구간과의 LSP 거리를 측정한다.

측정된 거리 값이 문턱값을 넘는 경우는 음소의 변화가 일어난 것으로 판정하고 이전에 음소가 변화된 구간에서 진행된 시간을 계산한다.

$$SPR = \frac{Fs}{VST(n) - VST(n-1)} \tag{10}$$

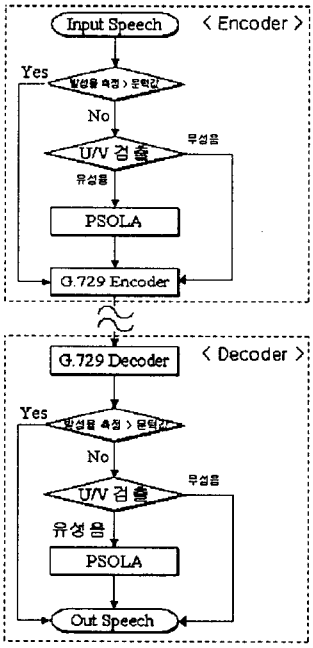

그림 7. 제안한 알고리즘의 블록다이어그램

5. 실험 및 결과

본 논문에서 제안한 방법을 시뮬레이션하기 위해 IBM-PC/Pentium-555 MHz에 마이크 입력이 가능한 16 비트 A/D변환기를 인터페이스하여 8 kHz의 표본화율로 16 비트 양자화하여 저장하였다. 시뮬레이션시 피치분석 프레임 단위를 240 표본으로 사용하였으며, 부프레임 길이는 60 표본으로 하였다. 피치주기 단위로 부호화 하였다. 처리결과의 성능을 측정하기 위해 다음의 대표적인 문장을 연령층이 다양한 남녀 5 명의 화자가 각 5 번씩 발성하여 시료로 사용하였다. 시료는 두드러진 피크를 가지지 않고 잡음이 30 dB를 가진 방에서 녹음하였다.

발성 1) "인수네 꼬마는 천재소년을 좋아한다."
발성 2) "창공을 날으는 인간의 도전은 끝이 없다."
발성 3) "예수님께서 천지창조의 교훈을 말씀하셨다."
발성 4) 일기예보 아나운서 음성시료

원 음성은 일반인을 이용하여 채취하였는데 그 이유는 훈련된 발화자보다 일반 사용자의 음성을 정확히 반영한다고 볼 수 있기 때문이다.

표 1. 지속시간 변경을 적용한 알고리즘의 전송률 비교

	G.723.1 (5.3kbps)	Proposed Method	Degradation bs
발성 1	5.251	2.6529	2.0478
발성 2	4.656	2.0486	2.6074
발성 3	5.044	3.7241	2.6734
발성 4	4.999	2.9954	2.0036

표 2. PSOAL기법을 적용한 알고리즘의 전송률 비교

	G.723.1 (5.3kbps)	Proposed Method	Degradation bps
발성 1	5.251	3.991	1.260
발성 2	4.656	2.447	1.209
발성 3	5.044	3.724	1.320
발성 4	4.999	3.670	1.329

표 3. 발성률을 적용한 알고리즘의 전송률 비교

	G.723.1 (5.3kbps)	Proposed Method	Degradation bps
발성 1	5.299	5.019	0.28
발성 2	4.759	4.457	0.302
발성 3	5.245	4.944	0.301
발성 4	5.019	4.748	0.271

표 4. PSOLA 압축률에 따른 MOS

압축률 (부호화된 음성/전체 음성) * 100	MOS
45.6%	4.1
38.8%	3.9
30.4%	3.7
23.9%	3.1

그리고 주관적 음질 평가를 하기 위해 MOS(Mean Opinion Score)를 사용하였으며 P.81을 준수한 MNRU Ver.2.0을 사용하였다. 제안한 방법을 C-언어로 구현하여 5.3 kbps ACELP (ITU-T 표준안 G.723.1) 보코더에 적용하였다.

6. 결 론

G.723.1 보코더는 인터넷 폰이나 화상회의, voice mail system, voice-pager 등에 응용이 가능하며 현재 상용버전으로 나와 사용되고 있다. 이 중 G.723.1은 5.3/6.3 kbps의 이중 전송률을 갖는 구조로 되어 있다. 최적의 전송 환경을 위하여 두 개의 전송률을 사용하기 때문에 다른 보코더 표준안들에 비해서 더욱 응용성이 높다. 그러나 G.723.1 역시 음성신호를 성분 분리하여 합성하는 방식인 CELP 보코더 계열의 합성에 의한 분석방법을 사용하기 때문에 많은 계산량으로 인한 처리 시간의 소모를 피할 수 없다는 문제점을 갖고 있다. G.723.1은 두 개의 서로 다른 보코더를 포함하고 있어 DSP칩으로 구현시 많은 내부 메모리와 계산량을 필요로 한다. 본 논문에서는 음성 데이터를 G.723.1 보코더 입력하기 전에 전처리단을 이용하여 전송률을 감소하고자 한다. 전처리단에 응용되는 기술은 기존의 파형 압축방법과는 전혀 다른 피치단위로 파형을 부호화하여 범용칩으로도 합성이 가능한 방법이다. 일반적으로 합성이나 인식에서 사용되는 파라미터인 지속시간, 발성율과 PSOLA 기법을 사용하였다. 입력음성이 보코더단에 들어가기 전 먼저 전처리단을 구성하여, 지속시간 변경을 통하여 입력음성의 지속시간을 1/2배 줄여 처리하는 방법과, 입력 음성의 발성률에 따른 달리 처리하는 방법, 그리고 NAMDF를 사용하여 음성의 주기를 찾아 압축한 다음 PSOLA 기법을 사용하여 다시 합성하는 방법을 적용하였다. 이 세 가지의 압축방법으로 전송률이 각각 46%, 5.6%, 31%씩 감소한 것을 알 수 있었다. 하지만, 지속시간의 변경법의 경우는 유성음에 대해서만 압축을 수행하고 있으므로, 차후 연구에는 무성음 및 묵음에 대해서도 압축을 수행할 수 있도록 하여, 좀 더 높은 압축률을 얻을 수 있어야 한다는 향후과제를 남기고 있다. 여기서 제안되는 음성부호화법의 특징은 알로리즘이 매우 간단하는 특징을 들 수 있는데, 이는 음성부호화법을 이용하여 상품화하려는 분야에 다음과 같은 알고리즘들을 좀더 보완하여 이용한다면, 저가의 범용칩을 이용해 상품화할 수 있으므로 대외경쟁력을 갖을 수 있을 것으로 예상된다.

참 고 문 헌

[1] ITU-T Recommendation G.723.1. March, 1996.

[2] Bae, M. J., Kim, D. S., Jeon, H. Y. and Ann, S. G. "On a new predictor for the waveform coding of speech signal by using the dual autocorrelation and the sigma-delta technique." *IEEE Proc. of ISCAS'94,* vol.6, No.3, pp.261- 264. June, 1994.

[3] 배명진 외 1인. "On Detecting the Steady State Segments of Speech Waveform by using the Normalized AMDF." 대한전자공학회지, Vol.14, No.1, pp.600-603. Jun., 1991.

[4] Kondoz, A.M. "*Digital Speech.*" John Wiley & Sons Ltd., Baffins Lane. Chichester, England. 1994.

[5] Rabiner, L.R. and Schafer, R.W. "*Digital Processing of Speech Signals.*" Prentice-Hall. Englewood Cliffs, New Jersey, 1978.

[6] Chapentier, F., Stella, M. G. "Diphone Synthesis Using Overlap-add Technique for Speech Waveforms Concatination." ICASSP 86, pp.2015-2018. 1986.

[7] Moulines, E. and Charpentier, F. "Pitch- synchronous waveform processing techniques for test to-speech synthesis using diphones." Speech Comm., Vol. 9 No. r, pp.453-467. 1990.

[8] Jayant, N.S. and Noll, P. *Digital Coding of Waveforms-Principles and Applicants to Speech and Video.* pp.220-221. Prentice-Hall. 1978.

접수일자: 2003. 11. 15.
게재결정: 2003. 12. 15.

▲ 장경아
서울특별시 동작구 상도5동 1-1 (우: 156-743)
숭실대학교 정보통신공학과 음성통신연구실
Tel: +82-2-824-0906
E-mail: kajang74@hotmail.com

▲ 민소연
서울특별시 동작구 상도5동 1-1 (우: 156-743)
숭실대학교 전자공학과 음성통신연구실
Tel: +82-2-824-0906
E-mail: pasternak@hanmail.net

▲ 배명진
서울특별시 동작구 상도5동 1-1 (우: 156-743)
숭실대학교 정보통신공학과 음성통신연구실
Tel: +82-2-820-0902
E-mail: mjbae@ssu.ac.kr