PSOLA 전처리과정을 이용한 G.723.1 보코더의 전송률 감소에 관한 연구

장 경 아, 조 성 현, 배 명 진
숭실대학교 정보통신공학과
156-743 서울시 동작구 상도동 1-1
kajang74@hotmail.com mjbae@saint.soongsil.ac.kr

On a Study of the Reduction of Bit Rate by the Preprocessing of PSOLA Coding Technique in the G. 723.1 Vocoder

KyungA Jang, SeongHeon Cho, Myung Jin Bae
Dept. Telecommunication Engr., SoongSil University
1-1 Sangdo-5Dong, Dongjak-Ku, Seoul 156-743, KOREA
kajang74@hotmail.com mjbae@saint.soongsil.ac.kr

* 이 논문은 정보통신진흥원 대학기초 인력사업 연구지원비에 의해 이루어졌습니다.

Abstract

In general, speech coding methods are classified into the following three categories: the waveform coding, the source coding and the hybrid coding. In this paper, First, the reference waveform is detected after searching the pitch period by NAMDF similarity and similarity between the reference waveform and the waveform each pitch period. It made a decision whether the waveform is compressed with the threshold of similarity. If the waveform is compressed only magnitude and pitch information is transmitted into the input of G.723.1 vocoder. Performing through the G.723.1 vocoder, the waveform is restored with the magnitude and pitch information by PSOLA synthesis method. The result of simulation with proposed algorithm has a 31% reduction of bit rate than the standard 5.3kbps G.723.1 ACELP vocoder.

Ⅰ. 서론

G.723.1 보코더는 인터넷 폰이나 화상회의, voice mail system, voice-pager 등에 응용이 가능하며 현재 상용버전으로 나와 사용되고 있다[2]. 이 중 그림 1-1.에 도시되어 있는 G.723.1은 5.3/6.3kbps의 이중 전송률을 갖는 구조로 되어있다[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[3]-[5]. G.723.1은 두개의 서로 다른 보코더를 포함하고 있어 DSP칩으로 구현시 많은 내부 메모리와 계산량을 필요로 한다. 논문에서는 G.723.1 5.3kbps ACELP를 기반으로 하여 음질을 유지하면서 전송률을 5kbps정도로 낮출 수 있는 새로운 부호화 방법을 제안한다. 본 논문에서는 음성 데이터를 G.723.1 보코더 입력하기 전에 전처리단을 이용하여 전송율을 감소하고자 한다. 전처리단에 응용되는 기술은 기존의 파형 압축방법과는 전혀 다른 피치단위로 파형을 부호화하여 범용칩으로도 합성이 가능한 방법이다.

Ⅱ. NAMDF에 의한 포만트 유사도 측정

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

$$NAMDF(d)=\frac{\sum_{n=1}^{N}|s(n)-s(n-d)|}{\sum_{n=1}^{N}|s(n)|+|s(n-d)|} \quad (2.1)$$

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

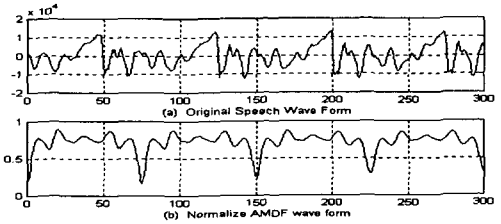

그림. 2-1 (a)음성파형 (b) NAMDF 파형

자기상관함수와 AMDF를 취했을 경우 영점 위치를 살펴보면 자기상관함수의 정확한 피크 값을 찾는 것이 AMDF의 피크 값을 찾는 것 보다 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 안의 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 알수 있다. 이러한 포만트의 정보는 한 피치주기 사이에 나타나는 피크의 수와 모양, 크기, 위치 등에 좌우된다. 따라서 포만트의 유사도를 측정하기 위하여 기준 피치와 인근 피치 주기내에 나타나는 피크들의 특성을 비교하였다. 한 주기안에 나타나는 피크들의 특성을 비교하기 위하여 기준피치와 인근피치 한 주기 파형에 대해 Cross NAMDF를 수행하였다.

$$NAMDF_{Cross}(d)=\frac{\sum_{n=1}^{N}|S_{ref}(n)-S_p(n-d)|}{\sum_{n=1}^{N}|S_{ref}(n)|+|S_p(n-d)|} \quad (2.2)$$

Cross NAMDF는 식 (2.2)과 같다. 여기서 S_{ref} 는 기준이 되는 피치주기의 파형이고 S_p는 p번째 주기의 파형이다. N은 윈도우 크기이고 S_{ref}와 S_p 길이중 작은값이다. d는 지연인자이다. 구해진 파형에 대한 면적은 식 (2.3)와 같이 구해진다.

$$A(p) = \frac{1}{N}\sum_{d=1}^{N} NAMDF_{Cross}(d) \quad (2.3)$$

여기서 $A(p)$는 p번째 파형의 Cross NAMDF 파형의 면적이다. 구해진 면적과 기준 피치주기의 NAMDF 파형의 면적을 비교하여 유사도를 측정한다. 유사도 측정은 식 (2.4)과 같다.

$$D(p) = \frac{|A_{ref}-A_p|}{A_{ref}} \times 100 \quad (\%) \quad (2.4)$$

A_{ref}는 기준파형의 NAMDF 파형의 면적이고, A_p는 식 (2.3)와 같이 구한 기준파형과 인근 파형의 Cross NAMDF 파형의 면적이다. D(p)는 p번째 파형의 포만트 유사도를 나타내며, 값이 작을수록 p번째 파형은 기준 파형과 유사하다.

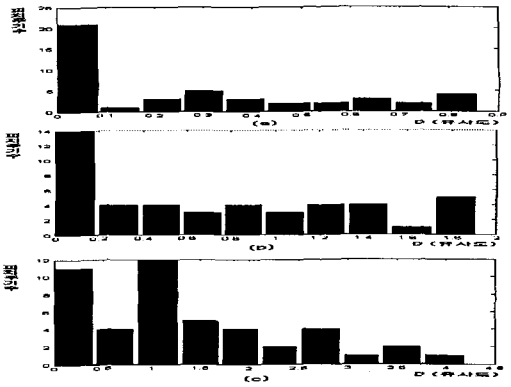

그림. 2-2 문턱값에 따른 피치주기의 압축률
첫막대(D=0) -> 전송되는 피치 주기 수
그외 (D>0) -> 압축되는 피치 주기 수
(a) D = 1를 문턱값으로 했을 때 (45.6%)
(b) D = 2를 문턱값으로 했을 때 (30.4%)
(c) D = 5를 문턱값으로 했을 때 (23.9%)

III. PSOLA 기법에 의한 음성합성

본 논문에서는 음성신호를 복원할 때 스팩트럼 왜곡률과 복잡성이 적은 PSOLA 방법이 적합하다[6]. 전송 또는 압축된 파형과 진폭정보와 피치정보를 이용하여

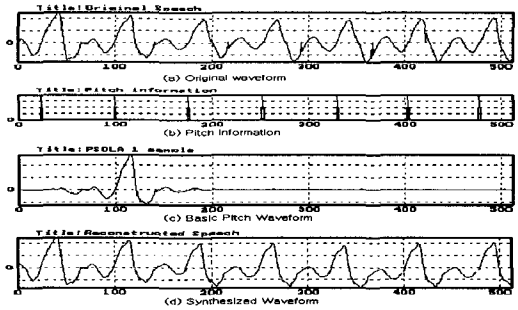

그림 3-1. 피치단위의 처리과정 예

PSOLA합성을 수행한다[7]. 그림 3-1은 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 3-2은 G.723.1 보코더에 입력하기 전 전처리과정을 나타내는 블록도이다. 송신단에서는 먼저 한 프레임에 대한 NAMDF법을 사용하여 피치를 0구한다. 피치는 그림 2-1 (b)에서 가장 먼저 0점에 가까워지는 Valley 까지의 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 3-3 (a)은 원래 음성 파형이고, (b)는 피치정보를 표시한 그림이고, (c)는 합성을 위한 한 주기 파형이며, (d)는 PSOLA 방법을 이용하여 합성한 파형이다.

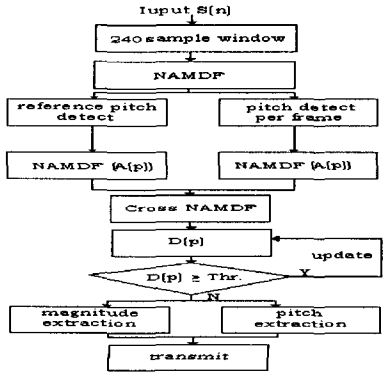

그림 3-2. 제안한 방법의 블록다이어그램

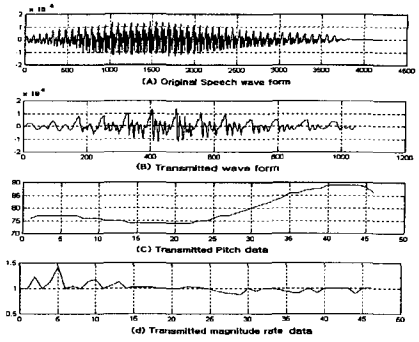

그림.3-3 '아' 음성에 대한 부호화
(A)음성파형 (B)전송되는 파형
(C)전송되는 피치정보 (C)전송되는 진폭(변화)정보

그 후 구해진 피치만큼의 파형을 기준 파형과 식(3.1)처럼 Cross NAMDF 수행하여 식(2.3)로 면적 A(p)와 구해진 면적과 기준면적으로 식(2.4)처럼 유사도 D(p)를 측정한다. 유사도가 문턱값 보다 작다면 압축하고 위와 같은 과정을 반복하고 만일 유사도가 문턱값 보다 크다면 그 주기를 기준 파형으로 하여 기준면적을 다시 구한 후 위와 같은 과정을 반복한다. 합성단에서는 전송된 파형과 피치정보, 진폭정보를 이용하여 PSOLA 방법으로 복원해낸다. 송신단에서 문턱값을 변화시킴으로써 압축율을 조정할 수 있다.

IV. 실험 및 결과

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

발성1) "인수네 꼬마는 천재소년을 좋아한다."
발성2) "창공을 날으는 인간의 도전은 끝이 없다."
발성3) "예수님께서 천지창조의 교훈을 말씀하셨다. "
발성4) 일기예보 아나운서 음성시료

원 음성은 일반인을 이용하여 채취하였는데 그 이유는 훈련된 발화자 보다 일반 사용자의 음성을 정확히 반영한다고 볼 수 있기 때문이다. 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 5.3kbps ACELP (ITU-T 표준안 G.723.1) 보코더에 적용하였다. 앞서 설명한 블록도와 같이 NAMDF를 이용하여 포만트 유사도 측정을 미리 측정된 기준파형과 유사도가 문턱값을 넘지 않은 경우에는 이에 대한 1비트 정보비트를 전성하므로써 압축하였고, 피치정보는 8kbps 표본 데이타에서 20에서 148이하 인가하므로 피치에 대해서 7비트 할당하여 수행하였다. 진폭정보는 Normalize한 후에 dynamic range를 측정하여 결정하고 초기값과 해상도에 대해서 부호화하였다. 합성시에는 앞서 전송된 비트정보를 가지고 진폭정보와 피치정보를 가지고 PSOLA방법으로 합성하였다. 표 4-1과 4-2는 G.723.1 보코더를 통과한 후 측정된 전송률과 복원된 파형에 대한 MOS test 결과를 나타내는 표이다. 전송률은 제안한 방법이 기존의 5.3kbps ACELP보다 31% 감소하였고 음질 열하는 자연성이 떨어지는 비음의 효과를 나타내었다.

표 4-1. 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 4-2. 기존 방법과 제안한 방법의 MOS 비교

	G.723.1 (5.3kbps)	Proposed Method
발성 1	3.84	3.2
발성 2	3.76	3.1
발성 3	3.8	3.2
발성 4	3.7	3.2

V. 결론

G.723.1 보코더는 인터넷 폰이나 화상회의, voice mail system, voice-pager 등에 응용이 가능하며 현재 상용버젼으로 나와 사용되고 있다. 이 중 G.723.1은 5.3/6.3kbps의 이중 전송률을 갖는 구조로 되어있다. 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 5.3kbps ACELP를 기반으로 하여 음질을 유지하면서 전송률을 5kbps정도로 낮출 수 있는 새로운 부호화 방법을 제안한다. 본 논문에서는 음성 데이터를 G.723.1 보코더 입력하기 전에 전처리단을 이용하여 전송율을 감소하고자 한다. 전처리단에 응용되는 기술은 기존의 파형 압축방법과는 전혀 다른 피치단위로 파형을 부호화하여 범용칩으로도 합성이 가능한 방법이다. 우선 NAMDF로 피치를 검색하여 기준 파형을 얻고 각 피치구간 별로 유사도를 측정한다. 유사도의 문턱값을 정하여 파형을 압축할 것인가를 결정한다. 압축할 경우에는 진폭과 피치정보만을 G.723.1 보코더 입력단에 전송한다. G.723.1보코더를 통과한 후 PSOLA 합성방식을 이용하여 파형을 복원하는 방법을 실험한 결과 기존의 5.3kbps ACELP보다 31%정도 감소하였다. 본 논문에서 제안한 음성부호화법은 유성음만 압축을 수행하고 있으나, 무성음 및 묵음에 대해서도 압축을 수행한다면, 좀더 높은 압축률을 얻을 수 있다. 본 논문에서 제안하는 음성부호화법의 특징은 알고리즘이 매우 간단하다는 특징이 있다. 따라서 음성부호화법을 이용하여 상품화하려는 분야에 본 논문에서 제안한 방법을 이용하여 음성데이터를 압축하여 전송하거나 저장할 경우 저가의 범용칩을 이용하여 상품화할 수 있으므로 대외 경쟁력을 가질 수 있다.

VI. 참고 논문

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

[2] M. J. Bae, D. S. Kim, H. Y. Jeon and S. G. Ann, "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] A.M. Kondoz *"Digital Speech"*, John Wiley & Sons Ltd, Baffins Lane, Chichester, England, 1994

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

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

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