비대칭 Weighting을 사용한 음성 피치변경법

함 명 규, 나 덕 수, 정 찬 중, 배 명 진
숭실대학교 정보통신공학과
156-743 서울시 동작구 상도동 1-1

On a Pitch Alteration of Speech Technique Using the Asymmetry Weighting

MyungKyu HAM, DuckSu NA, ChanJoong Jung, MyungJin BAE
Dept. Telecommunication Engr., SoongSil University
1-1 Sangdo-5Dong, Dongjak-Ku, Seoul 156-743, KOREA
mjbae@saint.soongsil.ac.kr

요 약

음성부호화의 주요목적은 대역 제한된 전송 대역폭에 전송을 하기위한 음성압축, 명료성과 자연성을 유지하는 고음질 음성합성, 그리고 처리 속도등의 요소에 따라 달라진다. 일반적으로 음성 부호화 방법은 파형 부호화법, 신호원 부호화법, 그리고 혼성 부호화법으로 나누어질 수 있다. 이러한 방법으로 전송되어진 음성은 다시 합성을 하게되는데, 이때 고음질을 유지할 수 있는 PSOLA법을 사용하였다.

본 논문에서 제안한 방법은 기존의 PSOLA 합성법에서 사용되어지는 Hanning 윈도우가 음성이 갖는 Glottal Wave Shape의 특성에 적합하지 않다는 것을 이용하여기존의 Hanning 윈도우가 아닌 비대칭성을 가진 새로운 형태의 비대칭 윈도우(Asymmetry Window)를 제안하였다. 비대칭 윈도우의 형태는 윈도우를 중심으로 왼쪽편은 기울기가 심하고, 오른쪽은 기울기가 완만하여 음성의 기울기에 적합한 웨이팅을 갖는 형태이다. 제안한 비대칭 윈도우를 사용하여 PSOLA 합성을 하였을 경우 SNR이 2~3dB 정도 향상되었음을 알 수 있다.

1. 서 론

음성합성은 저장된 음성시료들을 사용하여 기계가 음성을 발성하게 하는 기술이다. 이 때 기계에 의해서 발성되는 음성은 인간이 듣기에 명료하고 자연스러워야 한다. 기존에 사용되어지고 있는 합성 방식으로는 음성 신호의 특징 성분만을 파라미터들로 추출한 다음 이 성분들을 다시 합성하는 방식인 LPC 계열의 합성 방식이 대부분을 차지하고 있다. 이러한 방식은 음성 생성 모델을 근사화 한 것으로 모델링시에 생기는 오차에 의해서 고음질의 합성에서는 음질을 떨어뜨리게 된다.

PSOLA 기법은 음성파형을 그대로 이용하는 파형 부호화법을 사용하는데, 파형부호화법은 음성정보를 발성모델에 따라 분리하지 않고 파형 자체의 잉여성분만을 제거한 후 부호화하여 저장 또는 전송하고 필요에 따라 다시 원래 신호파형으로 합성하는 방법으로 LPC계열의 합성 방식보다 자연성이나 명료도가 뛰어나다[1]. 하지만 PSOLA 합성법에서는 음성파형 자체를 오버랩 시켜서 합성하므로 다른 형태의 단어등과 결합시 합성 음질이 저하된다는 단점을 가지게 된다. 첫째는, 윈도우가 피치와 정확히 맞지 않아서 생기는 것으로 이런 것은 위상성분이 왜곡되는 단점이 발생하게 된다. 둘째는, 서로 다른 단어들을 연결하는 경우에 피치가 맞지 않아서 발생되는 오차로 이것은, 이러한 경우에는 피치를 선형적으로 증가 또는 감소시켜서 문제를 해결할 수 가 있다. 셋째는, 합성시 spectral envelope의 차이로 발생하는 에러로 이것을 해결하기 위해서는 시간축에서 정확히 envelope을 맞추어 주는 방법이 있고, 다른 방법으로 주파수 영역에서 이를 보정하여 다시 시간축으로 변환하는 방법이 있다.

본 논문에서는 음성이 대칭이 아닌 비대칭성을 갖는다는 점에 착안하여 피치단위로 음성을 잘라내고, PSOLA에 적합한 윈도우를 적용하였다. 이 윈도우는 기존 윈도우와는 다른 웨이팅값을 갖는 비대칭 윈도우(Asymmetry Window)의 형태를 가지고 있다.

2. 기존의 피치변경법

파형 부호화법이나 혼성 부호화법에서의 피치 변경에 관한 연구는 거의 없었다. 대부분의 파형 부호화법이나 혼성 부호화 방법들은 종속되는 단어에 따라서 같은 단어를 서로 다른 데이터로 사용하였으며, 이는 데이터 베이스의 크기가 증가시킬 뿐만 아니라 또한 데이터 베이스 설계에 있어서도 문제가 된다. 이러한 한계들은 피치 변경법을 사용함으로서 극복될 수 있다. 지금까지 제안된 피치 변경법은 그 처리 영역에 따라 시간 영역법, 주파수 영역법, 시간-주파수 혼성 영역법으로 나누어 볼 수 있다. 시간 영역법에는 Multi-Pulse법, 피치 반분법 등이 있다. Caspers와 Atal은 MPLPC에서 펄스 사이에 영을 삽입하거나 삭제하는 방법을 제안했으나[3], MPLPC상의 펄스열은 피치와 포만트에 대한 상호연관을 갖고 있으므로 스펙트럼의 왜곡이 심하다. Varga와 Fallside는 LPC계수를 이용한 피치연장법을 제안했으나[8], 이 방법은 피치 주기를 줄이는 경우에 단지 파형의 일부분을 소거하고 평활화하는 방법을 사용하고 있기 때문에 스펙트럼의 왜곡이 많이 나타난다. 피치 반분법은 임의로 변경하려는 피치 주기의 2배 파형을 만든 후에 그 파형의 주기를 반분하는 피치 변경법이다. 그러나, 이 방법은 시간 영역에서만 수행되기 때문에 스펙트럼 왜곡이 발생하여 합성음의 명료성이 저하된다.

McAulay와 Quatieri는 주파수 영역에서 위상을 보존하는 피치변경법을 제안하였는데 이것은 입력된 음성에 대해 진폭 및 위상스펙트럼을 추출하여 별도로 처리하는 방법이다[5]. 진폭 스펙트럼에 대해서는 두드러진 스펙트럼 봉우리들을 추출한 다음에 이것을 피치변경율(ρ)만큼 인터폴레이션하여 진폭 스펙트럼의 피치를 변경시킨다. 위상 스펙트럼에 대해서는 시간 영역에서 구한 피치 개시 시간(pitch on-set time)에 해당하는 위상을 제거하고 나서 피치가 변경되었을 때의 새로운 피치 개시 시간의 위상을 더해 줌으로서 새로운 위상을 구성하게 된다. 이러한 방법은 파형의 꼴을 그대로 유지하기 때문에 프레임 단위로 분석 처리하는 통상의 처리법에서 인접 프레임간의 연결이 아주 용이해 진다는 장점이 있다. 그렇지만, 피치변경시에 피치 주기와는 별도로 피치의 개시 시간을 공급해 주어야 하고, 또한 진폭 스펙트럼 상에서 두드러진 봉우리 위주로 고조파의 인터폴레이션을 수행하기 때문에 스펙트럼의 왜곡이 높아진다는 단점이 있다.

시간-주파수 혼성법으로는 켑스트럼의 특징을 이용하여 켑스트럼값이 거의 영이 되는 부분에서 영값을 삽입하거나 삭제함으로써 피치를 변경하는 방법이 있다[1]. 그러나 이 방법 역시 위상의 보존이 어렵다는 문제점이 있다. Takagi와 Miyasaka가 제안한 시간-주파수 혼성법은 시간 영역에서 피치 변경을 하였을 때에 나타나는 스펙트럼 왜곡을 스펙트럼 영역상에서 LPC포락을 통해 수정하는 방법이다[7]. 이 방법은 LPC스펙트럼 포락이 갖는 극점에 치중된 시스템 전달 특성 때문에 모든 유성음을 만족하지 못한다는 한계성을 갖는다.

3. PSOLA 합성법

PSOLA 합성법은 그림 3-1과 같은 처리 절차를 따라 음성 신호를 합성하게된다.

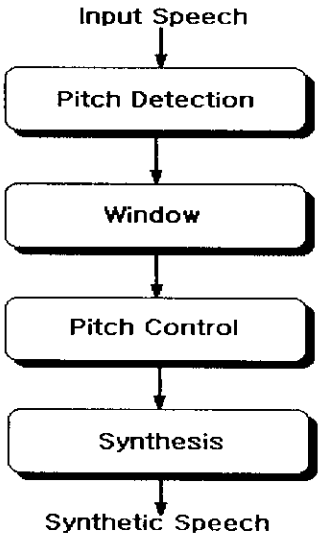

그림3-1. PSOLA의 블록다이어그램

첫째, 음성 신호를 피치 단위로 세그멘트(segment) 시키는 과정이다. 이 과정은 음성신호에서 피치 주기를 정확히 detection하여 피치 봉우리에 윈도우를 곱하여 피치 단위로 음성 신호를 잘라내는 과정이다. 이 때 사용하는 Hanning 윈도우는 보통 아래의 식(3.1)과 같은 함수식을 갖는 윈도우를 사용하였다. 여기서, N은 윈도우 사이즈를 나타낸다.

$$w(n)=\frac{1}{2}\left\{1-\cos\left(2\pi\frac{n}{N-1}\right)\right\},\ 0\le n\le N-1;\quad (3.1)$$

둘째, 피치 주기를 증가시킬 것인지 감소시킬 것인지를 조절하는 단계이다. 즉 피치 주기를 증가시킬 경우에는 위에서 세그멘트시킨 피치들이 가지고 있는 원 위치보다 간격을 조금씩 증가시켜서, 피치 간격을 넓힘으로써 피치 주기를 증가시킬 수 있다. 반대로 피치 주기를 감소시킬 경우에는 원 피치 주기의 간격을 좁힘으로써 피치의 주기를 감소시킬 수 있다.

셋째, 피치 주기를 단위로 합성을 하는 과정이다. 합성시에는 위의 두번째 단계에서 얻은 피치 주기 조절된 세그멘트 신호들의 재배치를 통해 합성하게 된다. 여기서 세그멘트 사이에 오버랩(overlap) 되는 부분들은 더하게 된다. 이렇게 하여 피치 주기가 변경된 PSOLA 합성 음성을 얻을 수 있다. 여기서, 단순히 비대칭성의 특성을 갖는 음성신호에 대해서 대칭성의 윈도우를 사용하게 되면, spectral envelope이 발생하게 된다. 따라서, 비대칭성의 형태를 가진 윈도우가 필요하게 된다.

4. 실험 및 결과

시뮬레이션을 수행하기 위한 장비와 음성시료는 다음과 같다.

- IBM-PC/Pentium MMX-200
- 16비트 A/D변환기를 인터페이스
- 3명의 남성화자와 2명의 여성화자
- 11kHz의 표본화율
- 16비트 양자화

발성1: **/인수네 꼬마는 천재소년을 좋아한다./**
발성2: **/예수님께서 천지창조의 교훈을 말씀하셨다./**
발성3: **/숭실대 정보통신과 음성통신연구팀이다./**
발성4: **/창공을 헤쳐나가는 인간의 도전은 끝이 없다./**
발성5: **/MAY I HELP YOU?/**

그림 4-1은 피치단위의 처리과정을 나타낸 것이다. 그림 (a)는 원신호의 음성신호를 나타내었고, (b)는 음성신호에서 피치를 찾아 표현한 것이다. 그리고, (c)는 하나의 기준 피치에 본 논문에서 제안한 비대칭 윈도우를 나타낸 것이고, (d)는 Hanning 윈도우를 나타낸 것이다. 또한 (e)는 하나의 피치에 대해서 비대칭 윈도우를 씌운것이다. 시뮬레이션을 수행하기 위해 한 프레임의 길이를 256표본으로 사용하였다. 처리 과정은 다음과 같다. 먼저 면적비교법을 사용하여 한 피치 구간의 음성표본을 피치단위로 자른 다음 기준 피치 파형으로 저장한다. 그리고 이 음성신호를 다시 합성하게 된다. 여기서 기준피치파형을 저장 할 때 쓰는 윈도우는 윈도우가 1이 되는 peak점을 기준으로 왼쪽 부분의 기울기가 크고, 오른쪽 부분의 기울기가 낮은 형태의 윈도우를 취하였다. 이것은 음성이 갖는 특성이 Glottal Wave Shape에 의해서 피치에서부터 서서히 감소하는 특성이 존재하기 때문에 윈도우의 기울기를 오른쪽은 서서히 감소시키고, 윈도우의 왼쪽 부분은 음성신호의 에너지가 서서히 감소하다가 피치 부분에서 다시 커진다는 특성에 의해서 기울기가 큰 윈도우를 사용하게되었다. 비대칭 윈도우의 함수식은 다음과 같다.

$$w(n)=\begin{cases}\frac{1}{2}\left\{1-\cos\left(2\pi\frac{n}{N-1}\right)\right\}, & 0\le n\le\frac{N-1}{3}\\ \frac{1}{2}\left\{1-\cos\left(\pi\frac{n}{N-1}\right)\right\}, & \frac{N-1}{3}\le n\le N-1;\end{cases}\quad (4.1)$$

식(4.1)과 같이 비대칭 윈도우는 왼쪽 33%의 기울기가 크고, 오른쪽 67%의 기울기가 완만한 형태를 갖는다.

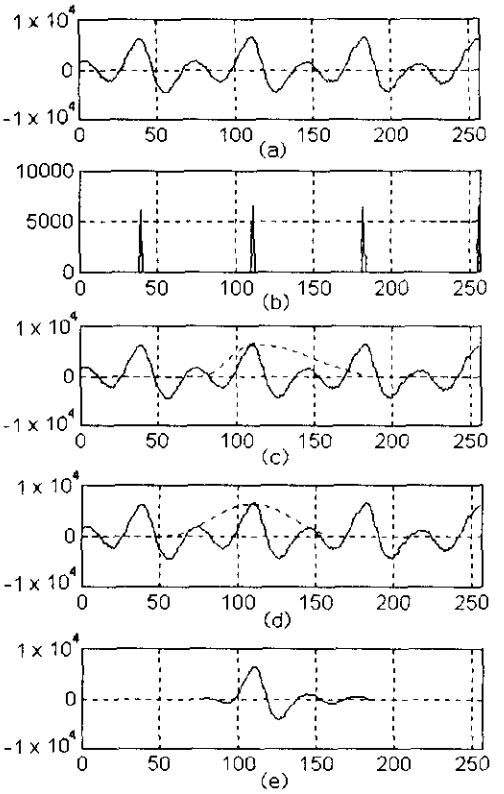

그림 4-1. 피치단위의 처리과정 예
(a) 원음성 신호, (b) 피치
(c) 비대칭 윈도우, (d) Hanning 윈도우
(e) 비대칭 윈도우를 사용하여 얻은 음성 신호.

이렇게 하여 얻은 합성 음성의 결과를 표 5-1에 나타내었다. 표 5-1에서 볼 수 있듯이 70%로 피치를 변경한 결과 약 13.40의 SNR을 얻을 수 있었고 130%로 피치를 변경하였을 경우 약 14.40의 SNR를 얻을 수 있었다.

표.5-1 피치 변경율에 따른 SNR(화자1)

	피치 변경율	
	70%	130%
기존 방법	10.78	11.35
제안한 방법	13.80	14.20
편 차	3.02	2.85

표.5-2 피치 변경율에 따른 SNR(화자2)

	변경율	
	70%	130%
기존 방법	11.06	11.56
제안한 방법	13.75	13.97
편 차	2.69	2.31

5. 결 론

음성신호를 성분별로 모델링하여 분리된 신호를 처리하는 신호원 부호화법은 모델링 자체에서 에러가 발생한다는 단점을 극복하기가 힘들다는 단점을 갖는다. 그러나, 신호원 부호화법은 데이터 양이 작다는 장점을 가지게 된다. 이와 반대로, 음성 신호를 분리하여 내지않고 파형 자체를 부호화하는 방법인 파형부호화법은 파형 자체를 사용하기 때문에 자연성과 명료성이 뛰어나다는 장점을 갖지만 데이터의 양이 상당히 많아진다는 단점을 가지게 된다. 이를 극복하기 위한 방법으로 고음질의 파형부호화를 사용하면서 데이터의 양을 줄일 수 있는 방법인 PSOLA 방법을 사용하게 되었다. 그러나, PSOLA 방법을 사용하여 합성을 하게되면 단어 연결시 spectral envelope이 맞지 않아 왜곡이 발생하게 된다. 이를 보완하기 위한 방법으로 음성의 비대칭성과 같은 특성을 갖는 비대칭 윈도우(Asymmetry Window)를 사용하여 spectral envelope왜곡을 최소화 하였다. 이 방법을 사용하여 피치를 70%로 줄인 경우와 130%로 피치 주기를 늘린 경우에 대해 실험한 결과 2-3dB정도의 SNR이 향상되었다.

5. 참고문헌

[1] N. S. Jayant and P. Noll, Digital Coding of Waveforms-Princioles 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] A. varga and F. Fallside, "A Technique for Using Multipulse Linear Predictive Speech Synthesis in Text-to-speech Type System", IEEE signal processing. Vol.ASSP-35, No.4, pp.586-587, APRIL 1987.

[7] M. BAE, "On the Pitch Alteration Methods for a High Quality Speech Synthesis", J., Acoust., Soc., Korea, Vol.12, No.2, pp.66-77, April 1993.

[8] B.E. Caspers and B.S. Atal, "Changing Pitch and Duration in LPC Synthesised Speech using Multipulse Excitation", J. Acoust. Soc. Amer., Vol.73, No.1, pp.55, Spring, 1983.

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

[10] 김상훈 외 2명, "Application of TD-PSOLA to Korean Text-to-Speech Conversion", SCAS-10권 1호 pp.291-294, 1993

[11] 임의택 외 2명, "A Study on the Improving Synthetic Speech Quality of TD-PSOLA TTS System", KSCSP '96 13권 1호 pp.292-295, 1996