Variable LPF에 의한 피치검출

(*)백금란 (**)배성근 (*)김대식 (*)배명진
(*)숭실대학교 정보통신학과 (**)건국대학교 전자공학과

(The Pitch Detection Using Variable LPF)

(*)Guemran BAEK (**)Seonggyun BAE (*)Daesik KIM (*)Myungjin BAE
(*)Soongsil University (**)Kunkook University

* 본 논문은 1995년도 KT 연구기초 연구지원금에 의하여 수행되었습니다. *

ABSTRACT

In speech signal processing, it is necessary to detect exactly the pitch. The algorithms of pitch extraction which have been proposed until now are difficult to detect pitches over wide range speech signals. Thus we propose a new algorithm which uses the G-peak extraction to do it. It is the method that finds the most MZI(maximum zero-crossing interval) at each frame and convolve it with speech signal ; this is the same with passing speech signals to variable LPF. Finally we obtained the pitch, improve the accuracy of pitch detection and extract it with the high speed.

Ⅰ. 서 론

음성인식, 합성 및 분석과 같은 음성신호처리 분야에 있어서 기본주파수 즉, 피치를 정확히 검출하는 것은 중요하다. 만일 음성신호의 기본주파수를 정확히 검출할 수 있다면 음성인식에 있어서 화자에 따른 영향을 줄일 수 있기 때문에 인식의 정확도를 높일 수 있고, 음성합성시 자연성과 개성을 쉽게 변경하거나 유지할 수 있다. 또한 분석시 피치에 동기시켜 분석하면 성문의 영향이 제거된 정확한 성도 파라미터를 얻을 수 있다.

이러한 피치검출의 중요성 때문에 피치검출에 대한 알고리듬이 다양하게 제안되었는데 그것은 시간영역법, 주파수영역법, 시간-주파수영역법으로 구분할 수 있다. 시간영역 검출법은 파형의 주기성을 강조한 후에 결정논리에 의해 피치를 찾는 방법으로 병렬처리법, AMDF법, ACM법 등이 있다. 이러한 방법은 보통 시간영역에서 수행되므로 영역과 변환이 불필요하고, 합, 차, 비교논리 등 간단한 연산만 필요하다. 그러나 음소가 천이구간에 걸쳐 있는 경우에는 프레임 내의 레벨변화가 심하고 피치의 주기가 변동하고 있기 때문에 피치검출에 어려움이 따르게 된다. 특히 잡음이 섞인 음성의 경우에는 피치검출을 위한 결정논리가 복잡해져서 검출 오류가 증가되는 단점이 있다.

주파수영역 피치검출법은 음성 스펙트럼의 고조파 간격을 측정하여 유성음의 기본주파수를 검출하는 방법으로 고조파분석법, Lifter법, Comb-filtering법 등이 제안되어져 있다. 일반적으로 스펙트럼은 한 프레임(20~40ms) 단위로 구해지므로, 이 구간에서 음소의 천이나 변동이 일어나거나 배경잡음이 발생하여도 평균화되므로 그 영향을 적게 받는다. 그러나 처리 과정상 주파수영역으로의 변환과정이 필요하므로 계산이 복잡하며, 기본주파수의 정밀성을 높이기 위해 FFT의 포인트수를 늘리면 그만큼 처리시간이 길어진다.

시간-주파수 혼성영역법은 시간영역법의 계산시간 절감과 피치의 정밀성, 그리고 주파수영역법의 배경잡음이나 음소 변화에 대해서도 피치를 정확히 구할 수 있는 장점만을 취한 것이다. 이러한 방법으로는 Cepstrum법, 스펙트럼비교법 등이 있고, 이 방법은 시간과 주파수영역을 왕복할 때 오차가 가중되어 나타나므로 피치추출의 영향을 줄일 수 있고, 또한 시간과 주파수영역을 동시에 적용하기 때문에 계산과정이 복잡하다는 단점이 있다.

따라서 본 논문에서는 위에서 열거한 문제점들 중 처리과정의 복잡성을 해결하고 측정의 정확도를 높일 수 있는 방법으로 가변 LPF에 의해 검출된 G-peak를 이용하여 피치를 검출하는 알고리듬을 새로이 제안하고자 한다. Ⅱ.장에서는 유성음 신호의 분석을 간단히 소개하였고, Ⅲ.장에서는 G-Peak의 정의와 Variable-LPF에 의해 검출된 G-peak를 이용한 피치검출법을 제안하였으며, Ⅳ. Ⅴ.장에서는 실험 및 결과를 평가한 후에 결론짓게 된다.

Ⅱ. 유성음 신호의 분석

음성신호는 음성원에 따라 유성음, 무성음, 파열음으로 구

분할 수 있다. 무성음의 경우에는 불규칙 잡음생성기가 그 생성원이므로 주기성은 나타나지 않지만, 주로 3KHz 근방에서 공진 봉우리를 갖기 때문에 유성음에 비해 평균 영교차율이 크다. 유성음은 폐에서 올라온 공기가 성문을 통하여 배출될 때 생성되므로 성대의 공진을 수반한다. 그리고 성도에서의 공명으로 인하여 아래 그림 1처럼 에너지가 크고 준-주기적인 형태의 신호가 된다.

이를 주파수영역에서 살펴보면 그림 2와 같이 성도의 공명 봉우리에 음성신호의 기본주파수 F_0가 세세하게 나타나고 있다. 성도 공명 봉우리에 해당하는 주파수들을 포만트라하고 가장 낮은 주파수를 갖는 봉우리를 제 1 포만트 F_1이라 한다.

유성음에서는 F_1이 다른 포만트들 보다 에너지가 약 10dB 이상 높다. 때문에 이를 시간영역으로 표현하면 F_1의 영향이 주로 나타나며 한 피치구간에서 zero crossing interval(ZCI)의 역수는 $2F_1$의 주파수와 거의 같게 된다. 그리고 포만트들은 대역폭을 갖게 되므로 시간영역의 한 피치구간에서는 감쇄진동을 하게된다.

F_1이 주파수 영역에서 다른 포만트들 보다 훨씬 높은 에너지 봉우리를 갖기 때문에 F_1만을 고려하여 근사적인 방법으로 성도를 분석할 수 있다. 그림 2에서 처럼 F_1의 크기가 대역폭 내에서 코사인 봉우리를 갖는다고 하면 이에 의한 시간영역에서의 파형은 그림 2를 IFT(Inverse Fourier Transform)하면 된다 (여기서 위상특성은 zero라 가정한다).

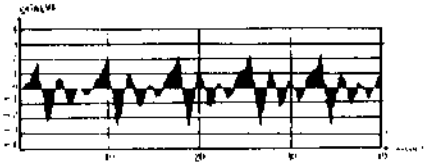

그림 1. 유성음 "아"에 대한 파형.
Fig 1. Waveform of voiced speech "a".

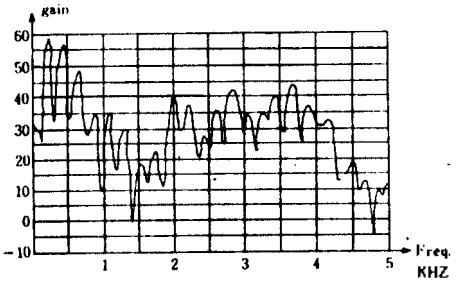

그림 2. 유성음 "이"에 대한 spectrum.
Fig 2. Spectrum for voiced speech "i".

$$h(t) = \int_{-\infty}^{\infty} F(f)\, e^{j2\pi ft} df$$

$$= \int_{-Bw/2}^{Bw/2} \cos\left(\frac{2\pi f}{B_w}\right) e^{j2\pi ft} df \cdot 2\cos\left((2\pi F_1 t) - \frac{\pi}{2}\right)$$

$$= \frac{4B_w}{\pi - 4\pi B_w^2} \cos(\pi B_w)\cos\left((2\pi F_1 t) - \frac{\pi}{2}\right) \cdots\cdots (2\text{-}1)$$

여기서 F_1는 제 1 포만트의 주파수이고 B_w는 F_1이 갖는 대역 폭이다.

식 (2-1)을 살펴보면 마지막 두 인자가 시간영역에서의 오실레이션을 결정하는데, 여기서 $F_1 >> B_w$라면 오실레이션은 F_1에만 의존하게 된다. 또한 식 (2-1)의 첫 항은 감쇄인자로 작용하는데, 이들기는 B_w에 관계됨을 알 수 있다. 한편, 유성음의 성대파를 발생하는 실제적인 방법을 그림 4에 보였다.

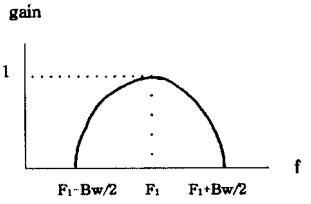

그림 3. 주파수 영역에서 제 1 포만트 근사분석
Fig 3. First formant approximation in frequency domain

임펄스열 제너레이터는 피치주기로 할당된 단위 임펄스의 시퀀스를 발생한다. 다음에 이 신호는 임펄스 응답이 성문파형 g(n)을 여기한다. g(n)의 형태는 단적으로 특정지을 수 없지만, Resonberg에 의해 합성 펄스파형 형태로 제시되었다[1].

$$g(n) = \frac{1}{2}\left(1 - \cos\left(\frac{\pi n}{N_1}\right)\right), \qquad 0 \le n \le N_1$$

$$= \cos\left(\pi \frac{n - N_1}{2N_2}\right), \qquad N_1 \le n \le N_1 + N_2$$

$$= 0, \qquad otherwise \quad \cdots\cdots (2\text{-}2)$$

식 (2-2)에서 $N_1 = 13$, $N_2 = 4$ 및 $n = 0.1msec$(표본주기)로 하였을 때 파형을 그림 4에 표시하였다.

$g(n)$이 유한 길이이므로 전극 모델이 바람직하게 되며, $G(Z) = z[g(n)]$에 대해 이극형 모델로 보통 모델링하고 있다. 그리고 방사의 효과는 $R(z) = R_0(1 - z^{-1})$로 나타낼 수 있으며, 이는 고역 필터로 동작하여 성도의 주된 공명효과를 강조 시키게 된다.

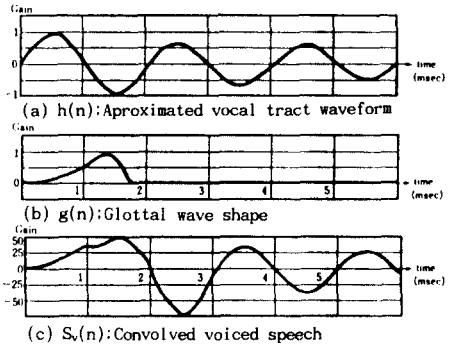

그림 4. 유성음의 근사분석

Fig 4. Approximation analysis for voiced speech.

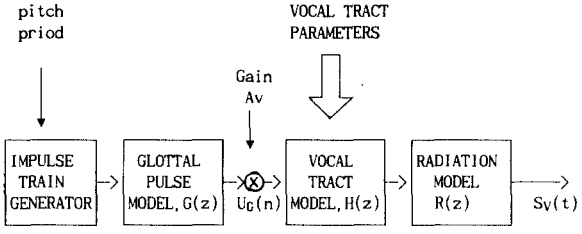

그림 5. 유성음의 생성모델

Fig 5. Speech production model for voiced signals

결국, 유성음 $S_v(n)$은 식 (2-1)과 식 (2-2)이 시간영역에서 콘벌루션된 것으로 나타난다.

$$S_v(n) \fallingdotseq h(n) * g(n) \quad \cdots\cdots\cdots\cdots \quad (2-3)$$

식 (2-3)은 그림 4(a)와 4(b)가 콘벌루션된 것을 나타낸다. 이로써 한 피치구간에서 처음 양의 봉우리가 다른 봉우리들 보다 강조되어 있음을 알 수 있다.

Ⅲ. G-peak에 의한 피치검출

그림 4(c)에서 유성음의 한 기본주파수 구간에서 처음의 양의 봉우리가 강조되는데 이것은 제 1 포만트 F_1이 대역폭을 가지고 있어서 감쇄진동을 하고 성문펄스가 그림 4(b)처럼 한쪽으로 치우쳐 있기 때문이다. 따라서 그림 4(c)에 있는 처음 봉우리가 성대성분과 F_1의 영향을 지배적으로 받게 된다. 여기서 처음 봉우리를 G-peak라 정의하는데 이는 성대성분이 지배적이라는 의미이다.

또한, G-peak가 식 (2-3)과 같이 성문펄스의 파형과 F_1의 포락선이 서로 컨벌루션된 신호의 첫 피크이기 때문에 성대의 영교차 간격은 성문펄스의 영교차 간격보다 길다. 그리고 제 1 포만트가 대역폭을 가지고 있기 때문에 그 파형은 감쇄진동을 하여 한 피치구간 내에서 첫 피크가 인근한 피크들에 비해서 진폭이 크게 된다.

시간영역에서 G-peak를 검출하려면 상위 포만트의 영향을 받기 때문에, 이를 감쇄하기 위해 유성음을 다음 식 (3-1)의 저역통과여파기(LPF)에 통과 시킨다.

$$S^*(n) = \frac{1}{N}\sum_{i=0}^{N-1} S(n-i) \quad \cdots\cdots\cdots\cdots \quad (3-1)$$

여기서 저역 통과여파기의 차단주파수 f_T는 식 (3-2)와 같기 때문에 차단대역을 식 (3-3)처럼 나타낸다 :

$$f_T = \frac{f_S}{N} \quad \cdots\cdots\cdots\cdots \quad (3-2)$$

또는, $$N = \frac{f_S}{f_T} \quad \cdots\cdots\cdots\cdots \quad (3-3)$$

이것은 매 프레임마다 G-peak의 폭이 변화하고 포만트의 성분비가 다르기 때문에 그것의 특성을 가변으로 해야한다. 즉, G-peak는 강조하고 포만트 성분은 감소시키려면 기본주파수와 포만트 주파수를 측정해서 저역 통과대역을 N을 가변해야 한다.

유성음 파형의 한 피치주기 내에서 영교차율은 제 1 포만트의 주기가 지배적이나 함께 나타나는 G-peak의 영교차 간격이 길다. 따라서, 한 프레임 내에서 가장 긴 영교차 구간을 선택하여 식 (3-1)의 저역 통과대역을 N으로 결정하게 된다.

먼저 유성음의 영교차점을 검출하고 각 영교차점의 간격을 식 (3-4)와 같이 계산한다.

$$ZCI(k) = Z_c(j+1) - Z_c(j) \quad \cdots\cdots\cdots \quad (3-4)$$

$$N = \max ZCI(k) \quad \cdots\cdots\cdots\cdots \quad (3-5)$$

(K는 1, 2, 3, 4, 5, 6 ·····)

여기서 $ZCI(k)$는 k번째 영교차구간이고, $Z_c(j)$는 j번째 영교차점이다. 식 (3-4)에서 계산된 ZCI들을 식 (3-5)와 같이 서로 비교하여 가장 큰 값을 구하면 이것이 어떤 한 프레임에서의 LPF의 차단대역을(N) 값으로 된다.

구해진 N을 적용하여 식 (3-1)를 2회 수행한다. 이것은 유성음에 대해 LPF를 두번 수행한 것과 같으므로 그림 6(b)와 같이 G-peak는 강조되고 상대적으로 포만트 성분은 감소된다. 또, 이 파형의 영이상 값을 취하면 G-peak만이 나타나는 파형을 얻을 수 있다.

가변 필터를 통과시킨 파형의 값 중 영이상의 값만을 취한 파형의 영교차점을 구한다. 그리고 식(3-6)에 있듯이 영교차점이 시작하는 점(N_S)과 끝나는 점(N_E)사이의 간격을 그 사이의 영교차율(PZCI)로 나누어 음성 파형의 피치를 검출한다 :

$$\text{Pitch} = \frac{N_S \quad N_E}{N_E} \qquad \cdots\cdots\cdots\cdots\cdots\cdots \quad (3\text{-}6)$$

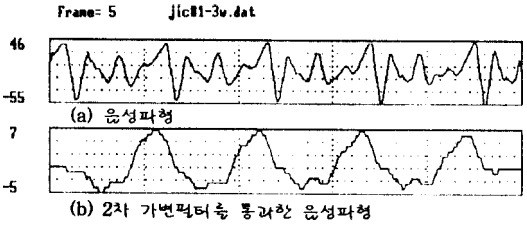

그림 6. 2차 가변 필터를 이용한 G-peak 검출.

Fig 6. G-peak detection using tow-variable LPF

Ⅳ. 실험 및 결과

이상의 과정을 컴퓨터 시뮬레이션하기 위하여 IBM PC/486에 마이크가 장치된 12-bit A/D변환기를 인터페이스 시키고, 아래의 발성음들을 8KHz의 샘플링 주파수로 표본화하여 저장한 다음 시뮬레이션의 시료로 사용하였다 :

발성 1) "인수네 꼬마는 천계소년을 좋아한다."
발성 2) "예수님께서 천지창조의 교훈을 말씀하셨다."
발성 3) "숭실대학교 음성신호처리 연구실이다."
발성 4) "공일이삼사오육칠팔구."

위의 각 음성시료에 대해 한 프레임의 길이를 256샘플로 하여 128샘플 단위로 오버랩하여 처리하였으며, 그림 7의 블럭도를 따라 순차적으로 수행하였다.

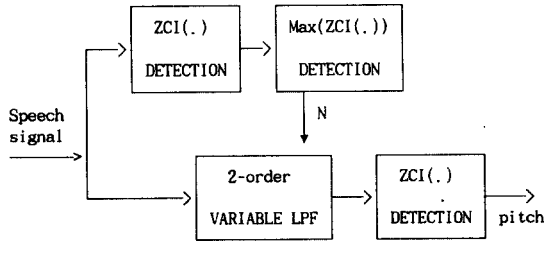

그림 7. 피치 검출에 대한 블럭도

Fig 7. Block diagram on pitch detection

성도 공진주파수의 포락선 위에 기본주파수가 겹쳐서 나타나는 유성음을 성대성분과 성도성분으로 분류한 후에 이들 두개의 성분을 컨벌루션시킬 때 근사된 유성음에선 첫번째 봉우리가 드러지게 된다. 즉, 성대성분이 지배적인 G-peak를 얻을 수 있는데 이것을 이용하여 본 실험에서는 위의 음성시료로 원래 유성음 파형의 매 프레임마다 영교차점을 찾은 후에 각각의 영교차점 간격을 구하여 이들 중 가장 큰 값의 간격을 구한다. 이때 얻어진 최대 영교차점 간격은 차단대역을이고 이것으로 두번의 컨벌루션을 시킨다. 여기서 두드러진 봉우리 G-peak를 찾을 수 있고 또한 이것으로 피치를 검출하였다.

그림 8의 (a)는 시간 영역에서의 음성파형을 나타내고, (b)와 (c)는 각각 영교차점과 영교차간격을 나타내며, (d)는 variable LPF에 의해 포만트 성분이 제거된 파형이다. 또한, (e)는 G-peak검출을 보여 주고 있다.

Ⅴ. 결 론

음성신호 처리영역에서 피치를 정확히 검출하는 것은 아주 중요하다. 피치가 정확히 검출될 수 만 있다면 음성인식, 합성 및 분석시 중요한 파라미터로 쓰일 수 있다. 즉 음성인식에 있어서 화자에 따른 영향을 줄일 수 있기 때문에 인식의 정확도를 높일 수 있고, 음성합성시 자연성과 개성을 쉽게 변경하거나 유지할 수 있다. 또한 분석시 피치에 동기시켜 분석하면 성문의 영향이 제거된 정확한 성도 파라미터를 얻을 수 있게 된다.

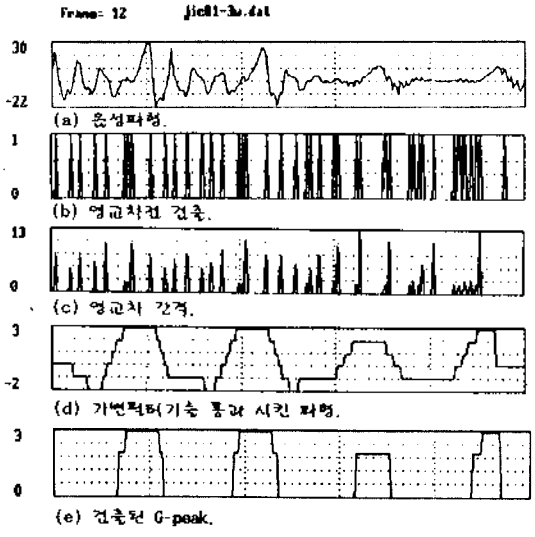

그림 8. 피치검출과정

Fig 8.The processing of pitch detection

따라서 본 논문에서는 시간영역에서의 유성음 파형에 대한 G-peak를 구하여 피치검출을 하는 알고리듬을 새로이 제안하였다. G-peak를 구하기 위해서는 LPF를 통과 시켜야 하는데 매 프레임 마다 G-peak와 폭과 포만트 성분비가 다르기 때문에 저역 통과여파기의 차단대역율을 가변하여야 하므로 가변 차단대역율을 구하여 이것으로 LPF를 두번 수행하면 G-peak는 강조되고 포만트 성분은 감소된다. 그리고 이것을 이용하여 피치를 검출하였다.

제안한 방법으로 시간영역에서 직접 피치를 검출하므로 처리과정 알고리듬이 간단해졌고 피치검출의 정확도가 향상되었다.

* 참고 문헌 *

[1] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech signals, Englewood Cliffs, Prentice-Hall, New Jersey, 1978.

[2] P. E. Papamichalis, Practical Speech Processing, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1987

[3] S. Seneff, "Real Time Harmonic Pitch Detection," IEEE Trans. Acoust. Speech, and Signal Processing, Vol. ASSP-26, pp.358-365, Aug. 1978.

[4] S. D. Stearns & R.A. David, Signal Processing Algorithms, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.

[5] M. Bae, and S. Ann, "Fundamental Frequency Estimation of Noise Corrupted Speech Signals Using the Spectrum Comparision", J.Acoust., Soc., Korea, Vol. 5, No. 3, 1989.

[6] E. Lee, C. Park, M. Bae, and S. Ann " The High speed Pitch Extraction of Speech Signals Using the Area Comparison Method" 대한전자공학회 Vol.22, No.2, pp.101-105, 1985.

[7] M. Bae, J. Rheem, and S. Ann "A Study on Energy Using G-peak from the Speech Production Model" 대한전자공학회 Vol.24, No.3, pp.381-386, 1987.

[8] Hans Werner Strube, "Determination of the instant of glottal closure from the speech wave", J. Acoust. Soc. Am.,Vol.5, No.5, pp.1625-1629, November 1974.

[9] M. Bae, I. Chung, and S. Ann, "The Extraction of Nasal Sound Using G-peak in Continued Speech" 대한전자공학회 Vol.24, No.2, pp.274-279, 1987.