전처리된 가변대역폭 LPF에 의한 피치검출법

°한진희, 장세현, 배 명진, 김명제(*), 김상룡(*)
숭실대학교 정보통신공학과, (*)삼성종합기술원 음성처리연구실

On a Pitch Detection using Low Pass Filter with Variable Bandwidth Preprocessed

J.H. Han, S.H. Jang, M.J. BAE, M.J. Kim(*), S.R. Kim(*)
Soongsil University, SAIT_SAMSUNG(*)

• 본 연구는 삼성종합기술원의 수탁과제 연구비 지원으로 이루어졌습니다.

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. In this paper, thus, we proposed a new pitch detection algorithm that uses a low pass filter with variable bandwidth. It is the method that preprocesses to find the first formant of speech signals by the FFT at each frame and detects the pitches for signals LPFed with the cutoff frequency according to the first formant. Applying the method, we obtained the pitch contours, improving the accuracy of pitch detection in some noise environments.

1. 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, 음성합성 시에 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 방법들이 다양하게 제안되었는데 그것은 시간영역법, 주파수영역법, 시간-주파수영역법으로 구분할 수 있다. 시간영역 검출법은 파형의 주기성을 강조한 후에 결정논리에 의해 피치를 검출하는 방법으로 병렬처리법, AMDF법, ACM법 등이 있다. 이러한 방법은 보통 시간영역에서 수행되므로 영역의 변환이 불필요하고, 합, 차, 비교논리 등 간단한 연산만 필요하다. 그러나, 음소가 천이구간에 걸쳐 있는 경우에는 프레임 내의 레벨변화가 심하고 피치주기가 변동하기 때문에 피치검출에 어려움이 따르게 된다. 특히 잡음이 섞인 음성의 경우에는 피치검출을 위한 결정논리가 복잡해져서 검출오류가 증가되는 단점이 있다[1][2].

주파수영역의 피치검출법은 음성 스펙트럼의 고조파 간격을 측정하여 유성음의 기본주파수를 검출하는 방법으로 고조파분석법[3], Lifter법, Comb-filtering법등이 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 필요함으로 계산이 복잡하며, 기본주파수의 정밀성을 높이기 위해 FFT의 포인터 수를 늘리면 그만큼 처리시간이 길어진다.

시간-주파수 혼성영역법은 시간영역법의 계산시간 절감과 피치의 정밀성, 그리고 주파수영역법의 배경잡음이나 음소변화에 대해서도 피치를 정확히 구할 수 있는 장점을 취한 것이다. 이러한 방법으로는 Cepstrum법, 스펙트럼비교법등이 있고, 이 방법은 시간과 주파수영역을 왕복할 때 오차가 가중되어 나타나므로 피치추출의 영향을 받을 수 있고, 또한 시간과 주파수영역을 동시에 적용하기 때문에 계산과정이 복잡하다는 단점이 있다[3][4].

따라서 본 논문에서는 상기에 열거한 문제점들 중 처리과정의 복잡성을 해결하고 측정의 정확도를 높일 수 있는 시간-주파수혼성형 피치검출법을 제안하고자 한다. 프레임 단위의 FFT에 의해 제 1 포만트를 검출하고 시간영역에서 이를 차단주파수로하는 가변대역폭 저역통과여파기에 통과된 신호를 통해 음성신호의 피치를 검출하는 방법이다. 이

검출법은 또한 한 피치구간에서 성문특성이 아주 지배적인 G-peak도 동시에 구할 수 있는 특징이 있다. 제 2 절에서는 유성음 신호의 분석을 간단히 소개하였다. 그리고 제 3 절에서는 가변 대역폭 LPF에 통과된 신호에서 피치 및 G-peak를 검출하는 방법을 나타내고, 4 절에서는 실험 및 결과를 평가한 후에 5장에서는 결론을 짓게 된다.

2. 유성음 신호의 분석

음성신호는 음성 여기원에 따라 유성음, 무성음, 혼합음으로 구분할 수 있다. 무성음의 경우에는 백색 가우시안 불규칙시퀀스가 그 여기원이므로 주기성은 나타나지 않지만, 주로 3kHz 근방에서 첫번째의 공진봉우리를 갖기 때문에 유성음에 비해 평균 영교차율이 크다. 유성음은 폐에서 올라온 공기가 성문을 통하여 배출될 때 진동되고, 성도에서의 공명으로 인하여 그림 2-1(a)처럼 에너지가 크고 준-주기적인 형태의 신호가 된다. 이를 주파수영역에서 살펴보면 그림 2-1(b)와 같이 성도의 공명봉우리에 음성신호의 기본주파수 F_0가 세세하게 나타나고 있다. 성도 공명 봉우리의 주파수들을 포만트라고 하고 가장 낮은 주파수의 봉우리를 제 1 포만트(F_1)라 한다.

일반적인 유성음 구간에서 F_1의 에너지봉우리는 다른 포만트들보다 10dB이상 높기 때문에 이를 시간영역의 파형으로 표현하면 F_1의 영향이 주로 나타난다. 한 피치구간에서 Zero Crossing Interval (ZCI)의 역수는 $2F_1$의 주파수와 거의 같게 된다. 그리고 포만트들은 대역폭을 갖게 되므로 시간영역 파형의 한 피치구간에서는 감쇄진동을 하게된다.

F_1이 주파수 영역에서 다른 포만트들보다 월등 높은 에너지 봉우리를 갖기 때문에 F_1만을 고려하여 근사적인 방법으로 성도를 분석할 수 있다. 그림 2-2에서 처럼 F_1의 크기가 대역폭내에서 코사인 봉우리를 갖는다고 하면 이에 의한 시간영역에서의 파형은 그림 2-2를 IFFT(Inverse Fourier Transform)하면 된다(여기서 위상특성은 zero라 가정한다). 여기서 F_1는 제 1 포만트의 주파수이고 Bw는 F_1이 갖는 대역폭이다.

$$h(t) = \int_{-\infty}^{\infty} F(f)e^{j2\pi ft}df$$
$$= \int_{-Bw/2}^{Bw/2} \cos\left(\frac{2\pi f}{Bw}\right)e^{j2\pi ft}df * 2\cos((2\pi F_1 t) - \frac{\pi}{2})$$
$$= \frac{4Bw}{\pi - 4\pi Bw^2}\cos((2\pi F_1 t) - \frac{\pi}{2}) \qquad (2\text{-}1)$$

식 (2-1)을 살펴보면 마지막 두 인자가 시간영역에서의 오실레이션을 결정하는데, 여기서 $F_1 >> Bw$라면 오실레이션은 F_1에만 의존하게 된다. 또한 식 (2-1)의 첫 항은 감쇄인자로 작용하는데, 기울기는 Bw에 관계됨을 알 수 있다.

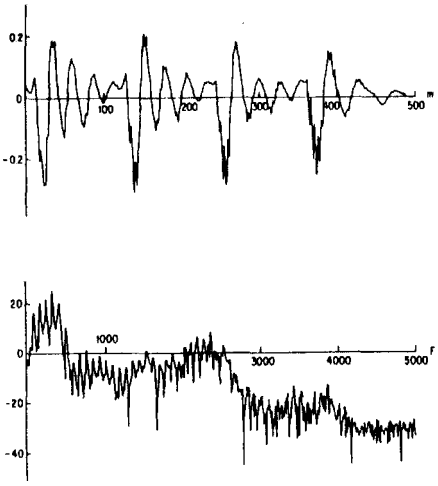

그림 2-1. 유성음의 파형과 spectrum :
(a) 유성음에 대한 파형,
(b) 유성음에 대한 spectrum.

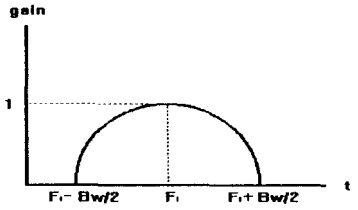

그림 2-2. 주파수 영역에서 제 1 포만트 근사분석

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

$$g(n) = \frac{1}{2}(1 - \cos(\pi\frac{n}{N_1})), \quad 0 \le n \le N_1$$
$$= \cos(\pi\frac{n-N_1}{2N_2}), \quad N_1 \le n \le N_1 + N_2$$
$$= 0, \text{ otherwise} \qquad (2\text{-}2)$$

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 성문이 갖는 전극필터의 효과를 일부 제거하게 된다.

결국, 유성음의 $s_v(n)$은 식 (2-1)과 식 (2-2)이 시간영역에서 컨벌루션된 것으로 나타난다.

$$s_v(n) = h(n) * g(n) \qquad (2\text{-}3)$$

따라서 한 피치구간에서 처음 양의 봉우리가 다른 봉우리들보다 두드러지게 나타나게 되고, 우리는 이 봉우리가 한 피치구간내에서 성문의 영향이 크게 나타나는 봉우리로 고려하여, G-Peak(glottal peak) 봉우리라고 하였다.

3. 전처리된 가변대역 LPF에 의한 피치검출

음성신호의 피치는 음성 파형의 반복되는 봉우리에서 봉우리까지 끝에서 끝까지로 정의된다. 파형의 봉우리 위주로 피치를 검출하는 경우에는 두드러진 봉우리가 존재하는 시간지연에 대해서만 자기 상관관계가 높게 존재한다. 반면, 파형의 골에 의해 피치를 검출하는 경우에는 두드러진 골이 존재하는 시간 지연에 대해서만 자기 상관관계가 높게 존재한다.

음성 파형에 대해 피치주기를 검출하려고 하면 성도 포만트에 따른 영향을 받게 된다. 성도포만트들은 문장을 구성하는 음소에 따라 변화하게 되고, 음소는 10ms 정도의 범위 내에서는 안정상태를 이룬다. 또한 피치검출시에 크게 영향을 주는 포만트들은 기본주파수에 근접한 제 1 포만트이고, 이 포만트와 에너지가 파형을 지배하기 때문에 이 성분을 적응적으로 제거 또는 억압시킬 필요가 있다.

음성 파형의 제 1 포만트성분을 억압시키는 간단한 방법은 저역통과여파기(LPF)에 음성신호를 통과시키는 것이다. 이 경우에 LPF의 차단주파수는 제 1 포만트주파수와 기본주파수의 중간 영역을 차지하는 것이 바람직하다. 또한 여성 또는 어린이 화자의 발성이거나 /이/나 비음 음소와 경우에는 음성의 기본주파수와 제 1 포만트주파수가 거의 일치하기 때문에 LPF의 차단특성은 임격한 것보다 완만한 기울기를 유지할 필요가 있다. 따라서 우리는 다음과 같이 주파수영역에서 sinc(.)함수의 구조를 갖는 LPF를 사용하였다:

$$s(n)' = \frac{1}{N}\sum_{i=0}^{N-1} s(n-i) \qquad ,10 <= N <= 200 \qquad (3\text{-}1)$$

여기서 차단주파수는 $f_T = fs/N$ 이고, s(n)은 음성신호이다. 통계적으로 피치주기는 25msec 이내에서 찾아지기 때문에 차단주파수를 결정하는 구간 N은 표본화주파수 fs=8KHz에서 200표본 이하이다. 또한 제 1 포만트의 주파수는 한국인의 경우에 800Hz 이하에 존재하기 때문에 구간 N은 10표본 이상이 된다.

LPF의 차단주파수에 대한 구간 N을 정확히 결정하기 위해서는 주어진 프레임의 음성 파형에 대해 제 1 포만트를 얻어야 한다. 한 프레임의 음성신호 s(n)을 주파수영역으로 변환하여 전폭스펙트럼을 살펴보면 제 1 포만트의 에너지가 여타의 포만트보다 높다. 따라서 전폭스펙트럼 M(K)에서 최대의 에너지를 이루는 주파수 K_m을 측정하면 근사적인 제 1 포만트의 주파수가 된다. 이때 피치검출에 적용할 LPF의 차단주파수 fc는 다음과 같이 적용한다:

$$fc = 0.9\ K_m \qquad (3\text{-}2)$$

여기서 차단주파수를 제 1 포만트의 0.9 배로한 것은 제 1 포만트의 주파수와 기본주파수가 거의 일치하는 경우를 고려한 것이다.

유성음 파형이 이 LPF에 통과되면 저역대에 에너지가 집중적인 성문 성분은 강조되고 상대적으로 포만트 성분들은 감소된다. 따라서 통과된 파형은 성문구조의 근사모양을 나타내게 되어, 영을 교차하는 위치를 측정하면 피치주기가 시작되는 시점이 된다.

피치주기를 검출하기 위한 결정논리로는 이 가변 통과대역폭을 갖는 LPF에 통과시킨 파형의 값에 대해 파형의 문턱값(또는 영값)의 교차점을 사용한다. 주어진 프레임에서 문턱값을 입상(rising) 교차점이 시작하는 점(Ns)과 끝나는 점 (Ne)사이의 간격과 그 사이의 입상교차율(Rising Threshold level Crossing Rate, RTCR)로 나누어 다음과 같이 그 프레임의 평균 피치주기를 검출한다:

$$PITCH(fr) = \frac{N_E - N_S}{RTCR} \qquad (3\text{-}3)$$

식 3-3에 의해 fr번째 프레임에서 검출된 PITCH(fr) 값은 피치의 존재 영역인 2.5-25ms이내에 있어야 한다. 또한 검출된 값이 그 프레임내의 개별 입상교차점간의 간격에 비해 10%이내에 존재하면 올바른 피치주기로 판정한다. 이러한 조건이 만족되지 않는 경우에는 무성마찰음, 무성파열음, 묵음 등의 구간으로 처리한다.

4. 실험 및 결과

이상의 과정을 컴퓨터 시뮬레이션하기 위하여 IBM PC/486 DX2(60)에 마이크가 부착된 16-비트 A/D변환기를 인터페이스시키고, 아래의 문장들을 남녀 각 3명에게 발성시키면서 8kHz의 표본화 주파수로 표본화하여 저장한 다음에 시뮬레이션의 시료로 사용하였다 :

발성 1) "인수네 꼬마는 천재소년을 좋아한다."
발성 2) "예수님께서 천지창조의 교훈을 말씀하셨다."
발성 3) "숭실대 정보통신과 음성통신 연구팀이다."
발성 4) "창공을 헤쳐나가는 인간의 도전은 끝이없다."
발성 5) "공일이삼사오육칠팔구."

위의 각 음성시료에 대해 한 프레임의 길이를 256샘플로 하여 128샘플 단위로 오버랩하여 피치검출을 수행하였다. 본 본문에서 제안한 피치 검출과정을 블럭도로 나타내면 그림 4-1과 같다. 주어진 프레임내의 음성 파형(예를 들면 그림 4-2(a))에 대해 LPF의 차단주파수를 결정하기 위해 FFT를 수행하여 진폭스펙트럼을 구하고, 최대의 에너지 값을 갖는 위치를 찾는다. 이 주파수를 식 3-2와 같이 LPF의 차단주파수로 사용하여 여파기에 통과시킨다. LPF에 통과된 신호(예를 들면 그림 4-2(b))에 대해 문턱 값을 임상교차하는 간격으로 피치주기를 판정하였다. 이와 동시에 원래 음성 파형에 대해 찾아진 임상교차의 위치를 중심으로 +-3 표본 주위에서 영교차점을 구하여 피치싯점(예를 들면 그림 4-2(c))을 측정하였다.

상기의 음성시료에 대해 가우시안 백색잡음을 신호의 에너지에 비례적으로 가미하면서 피치검출에 대한 조오율(gross error)을 측정하여 표 1에 제시하였다. 결과적으로 제안한 방법은 시간영역에서 직접 피치를 검출하므로 파형의 위상특성을 유지하게되어 피치싯점이 함께 검출되고, 처리과정에서 LPF를 수행하기 때문에 고주파 대역성분에 의한 영향은 억압된다. 또한 통과대역의 결정은 주파수영역에서 처리하기 때문에 배경잡음에도 강인한 특성을 나타내었다.

5. 결 론

음성신호 처리영역에서 피치를 정확히 검출하는 것은 아주 중요하다. 피치가 정확히 검출될 수만 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, 음성합성시에 자연성과 개성

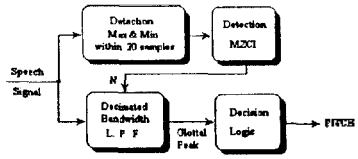

그림 4-1. 피치 검출에 대해 제안한 처리 블럭도

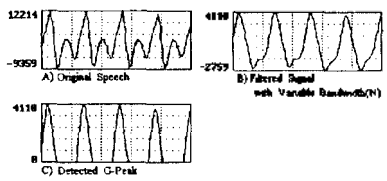

그림 4-2. 피치 검출과정의 결과 예시: (a)음성 파형, (b) 가변대역으로 저역여파된 신호, (c)검출된 피치싯점

표 1. 각 음성문장에 대한 gross 에러율

발 성	분 석 프레임 수	Gross Error Rates (%) clean	SNR 6dB	SNR 3dB	SNR 0dB
1	192	0.12	1.04	1.27	3.55
2	192	0.27	1.22	1.34	3.57
3	192	0.72	1.35	1.50	4.23
4	64	0.25	1.21	1.35	3.61
평 균		0.34	1.21	1.37	3.74

을 쉽게 변경하거나 유지할수 있다. 또한 분석시 피치에 동기시켜 분석하면 성문의 영향이 제거된 정확한 성도 파라미터를 얻을 수 있게 된다.

따라서 본 논문에서는 시간-주파수영역에서 얻을 수 있는 장점을 이용하는 혼성형 피치검출법을 새로이 제안하였다. 성문성분을 강조하고 성도성분을 억압하기 위해 프레임마다의 가변 통과대역을 갖는 LPF를 사용하였다. 그리고 LPF의 차단주파수는 결정이 간단한 주파수영역에서 취하였다.

제안한 방법은 시간영역에서 직접 피치를 검출하므로 파형의 위상특성을 유지하게되어 피치싯점이 함께 검출되고, 처리과정에서 LPF를 수행하기 때문에 고주파 대역성분에 의한 영향은 억압된다. 또한 통과대역의 결정은 주파수영역에서 처리하기 때문에 배경잡음에도 강인한 특성을 나타내었다.

6. 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 Comparison," J., Acoust., Soc., Korea, Vol. 8, No. 3, June 1989.

[6] E. Lee, C. Park, M. Bae, and S. Ann "The High Speed Pitch Extraction of Speech Signals Using the Area Comparison Method," KIEE, Korea, Vol. 22, No. 2, pp.13-17, March 1985.

[7] M. Bae, J. Rheem, and S. Ann "A Study on Energy Using G-peak from the Speech Production Model," KIEE, Korea, Vol. 24, No. 3, pp. 381-386, May 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," KIEE, Korea, Vol. 24, No. 2 pp. 274-279, March 1987.