G-Peak의 특성에 의한 피치시점검출

(*) 이해군 (**) 금 홍 (**) 배명진 (*)임운천
(*) 호서대학교 전자공학과 (**) 숭실대학교 정보통신공학과

(The Pitch Beginning Point Extraction Using Property of G-peak)

(*) Haegoon LEE (**) Hong KEUM (**) Myungjin BAE (*) Uncheon LIM
(*) Hoseo University (**) Soongsil University

* 이 연구는 1993년도 한국과학재단 연구비 지원에 의한 결과임.*
* 과 제 번 호 : 92-21-00-06 *

ABSTRACT

In this paper, a new pitch beginning point detection method by extracting the G-peak, is proposed. By the speech production model, the area of the first peak on a pitch interval of speech signals is emphasized. By using the above characteristics, this method have more advantages than the others for pitch beginning point detection. The defective decision caused by an impulsive noise is minimized and the pre-filtering is not necessary for this method, because the integration of signals takes place in the process.

I. 서 론

음성신호처리 분야에서 피치시점을 정확히 검출하는 것은 아주 중요하다. 만약 피치시점을 정확히 검출할 수 있다면, 음성분석시 피치에 동기시켜 분석할 수 있고, 인식시에는 성문의 영향이 제거된 정확한 성도 파라미터를 얻을 수 있으므로 인식의 정확도를 높일 수 있다. 또한, 합성시에는 여기원의 위상특성을 파악할 수 있으므로 개성이 강조된 합성음을 얻을 수 있다.

그동안 피치시점을 검출하기 위한 많은 연구가 진행되어 왔다. 그 중 Strube[8]는 성문 폐쇄의 순간(GCI)이 위치하는 음성파형의 공분산 행렬(covariance matrix)에 대한 log 행렬식(determinant)의 측정을 제안했다. 하지만 이 방법은 모든 모음신호들에 대해서 적용할 수 없다. 다시 말해서 실제 어떤 모음들에서는 임펄스와 큰 예측에러로 부터 GCI 주위에 발생하는 많은 잔여신호들 때문에 GCI(glottal closure instant)를 결정하기가 매우 어렵게 되며, 또한 처리시간이 많이 소요된다. Wong[9]은 음성파형의 M-점 창에서 p-pole 전체 선형 예측 에러 시퀀스를 해석하는 방식을 제안했다. 그러나, 이 방법은 성문의 폐쇄된 위상이 매우 짧은 구간을 가진 고주파나 호흡(breathy) 음성인 경우에서는 정확한 폐쇄위상을 얻는 것이 어렵다. 그리고, Veeneman은 EGG신호를 이용하는 방법을 제안하였다. 하지만 이 방법은 후두에 마이크로폰을 부착하여 직접적으로 성대의 움직임을 측정하여야 하며, 역 필터링 과정이 필요하다.

따라서 본 논문에서는 시간영역에서 직접 피치시점을 검출하면서 처리 알고리즘이 간단하고, 또한 배경잡음의 영향에도 강인하며, 음성신호의 위상성분은 그대로 유지하면서 정확히 피치시점을 검출하는 방법을 새로이 제안한다. 제안한 방법은 음성신호의 발성모델에 근거하여 성문특성이 지배적인 피크인 G-peak의 특성을 이용하여 피치시점을 검출하였다.

먼저, 제 II절에서는 음성생성 모델 대해서 알아본 뒤, G-peak를 정의하고, 제 III절에서는 본 논문에서 제안한 G-peak의 특성에 의한 피치시점 검출법에 대해 제안한 뒤, 실제 음성시료에 대해 처리한 결과를 검토하고 결론을 짓는다.

II. 음성생성 모델

음성신호를 음성신호의 생성모델[1] 측면에서 고려하면 그림 2-1에서 처럼 무성음의 경우에는 불규칙잡음생성기가 그 생성원이므로 주기성은 나타나지 않지만, 주로 3KHz 근방에서 공진 봉우리를 갖기 때문에 유성음에 비해서 평균 영교차율이 크다.[2]

유성음 생성 과정은 그림 2-2에서 보인 선형 시스템으로 모델화 된다[11]. 성대 v-v(golttal volume-velocity) 펄스열 g(n)은 성도 전달함수 H(Z)로 필터링되고, 이 구간 v-v의 결과는 v(n)이다. 즉, 유성음은 성문펄스가 그 생성원이며 성대의 진동에 따른 성도의 영향이 강조되어 나타나 일반적으로 진폭이 크고 준주기적인 성질(pitch)을 갖는다. 따라서 유성음의 에너지원은 glottal 성분이라고 볼 수 있다.

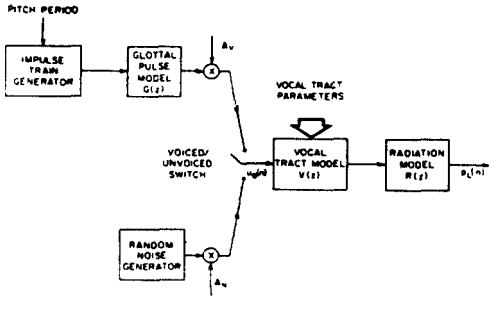

그림 2-1. 음성신호의 생성모델[1]
Fig. 2-1. Speech Production Model.[1]

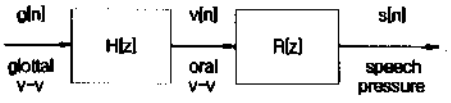

그림 2-2. 유성음 생성 모델.
Fig. 2-2. Voiced speech production model.

일반적으로 유성음은 성대의 진동과 성도의 공명 현상 때문에 시간영역에서 진폭이 크고 준주기적 성질을 갖는다. 이것을 주파수영역에서 살펴보면 그림 2-3에서 처럼 성도의 공진주파수의 포락선 위에 유성음의 기본주파수가 겹쳐서 나타난다. 그리고 제 1포먼트 주파수 F_1의 이득이 다른 포먼트보다 약 10dB이상 높으므로 F_1만의 포락선을 가지고 성대를 근사할 수 있다.[2,5]

그림 2-4에서와 같이 F1의 포락선이 대역폭내에서 cosine의 포락선을 갖는다고 하면 이것에 대한 시간영역에서의 파형은 이것을 역 퓨리어변환하면 얻을 수 있다.(여기서 위상특성은 제로라고 가정한다.)[5]

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

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

$$=\frac{4B_W}{\pi-4\pi B_W^2}\cdot\frac{1}{t^2}\cos(\pi B_W t)\cos(2\pi F_1 t-\frac{\pi}{2}) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2\text{-}1)$$

여기서 성문펄스모양은 Rosenberg에 의하여 합성된 파형을 적용할 수 있다.[1,4]

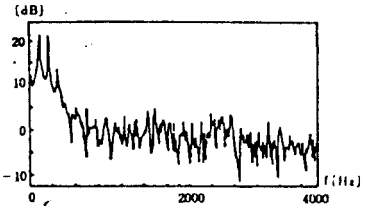

그림 2-3. "에"모음의 스펙트럼
Fig. 2-3. Spectrum of Vowel /ε/.

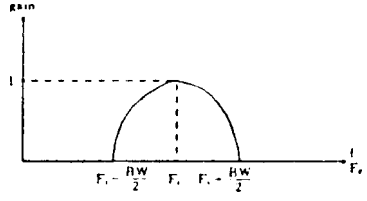

그림 2-4. 주파수영역에서 제 1포먼트의 근사분석[5]
Fig. 2-4. First Formant Approximation in Frequency Domain.[5]

$$g(n)=\begin{cases}\frac{1}{2}[1-\cos(\frac{\pi n}{N_1})] & ,0\le n\le N_1\\ \cos[\pi(n-N_1\frac{)}{2N_2}] & ,N_1\le n\le N_2\\ 0 & ,otherwise\end{cases} \quad \cdots (2\text{-}2)$$

따라서 음성신호 S(n)은 (2-1)식과 (2-2)식의 시간영역에서의 컨벌루션으로 근사될 수 있다.[1,2]

$$s(n)\fallingdotseq h(n)*g(n) \quad \cdots\cdots\cdots\cdots (2\text{-}3)$$

그림 2-5에 식(2-1), (2-2), (2-3)의 파형을 보였으며 특히 그림 2-5의 (c)에서 보면 유성음의 한 기본주파수 구간에서 처음의 양의 피크가 강조됨을 알 수 있다.[5] 이것은 제 1 포먼트 F_1이 대역폭을 가지고 있어서 감쇄진동을 하고 성문펄스가 그림 2-5의 (b)처럼 positive 쪽으로 치우쳐 있기 때문이다.[2] 따라서 그림 2-5의 (c)에서 처음의 피크가 glottal 성분과 F_1의 영향을 지배적으로 받는다 할 수 있다. 여기서 처음의 피크를 G-peak라고 정의하고, 나머지 피크들을 Side-peak라 하자. G-peak는 glottal 성분이 지배적인 피크라는 의미이다.

Ⅲ. G-Peak에 의한 피치시점 검출

연속음성신호에 대한 장시간(예를들면 1분이상) 전력밀도스펙트럼을 살펴보면 그림 3-1과 같이 500Hz근처에서 피크가 나타난다.[1,2] 즉 음성신호의 에너지는 주로 700Hz 이하에 몰려 있음을 알 수 있다. 이것은 성도의 제 1 포먼트 주파수 범위와 비슷하므로 F_1만의 포락선에 의한 성도의 근사가 타당함을 보여준다. 그리고 일반적으로 기본주파수 F_0가 50~500Hz 사이에 나타나므로 이것을 포함하는 G-peak의 특성이 유성음의 에너지원이 됨을 알 수 있다. 따라서 G-peak는 한 피치구간에서 음성신호의 성문특성과 F_1의 영향이 지배적으로 나타나는 피크이다.

또한 G-peak가 식 (2-5)와 같이 성문펄스와 파형과 F_1의 포락선이 서로 컨벌루션된 신호의 첫 피크이기 때문에 G-peak의 영교차간격은 성문펄스의 영교차간격 보다 길게 되고, 제 1포먼트가 대역폭을 가지고 있기 때문에 그 파형은 감쇄진동을 하여 한 피치구간 내에서 첫 피크가 인근한 피크들에 비해서 진폭이 크게 된다.

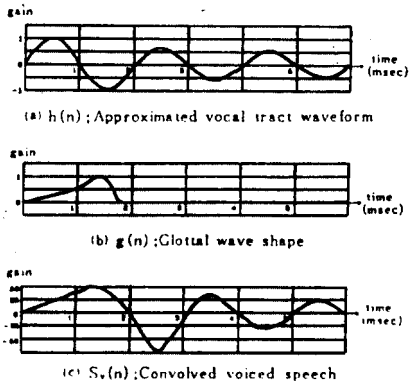

그림 2-5. 유성음의 근사분석[5]
Fig. 2-5. Approximation Analysis for Voiced Speech[5]

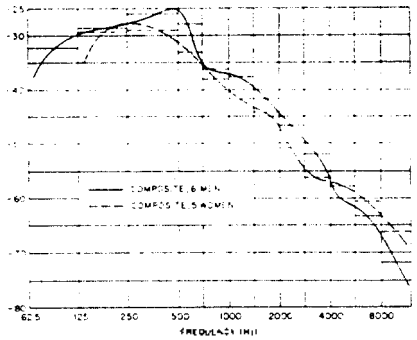

그림 3-1. 장시간 전력밀도스펙트럼[1]
Fig. 3-1. Long Time Power Density Spectrum[1].

그렇지만 음성신호에 대해 시간영역에서 G-peak를 검출하려고 하면 상위 포먼트들의 영향을 받아 결정논리가 복잡하게 된다. 따라서 음성을 파형을 다음 식 (3-1)의 저역통과필터에 통과시킨다.

$$S'(n-\frac{N}{2}) = \sum_{i=0}^{N-1} S(n-1) \quad \cdots\cdots \quad (3-1)$$

여기서 차단주파수 f_T 는 $\frac{f_s}{N}$ (또는 $N = \frac{f_s}{f_T}$)의 관계가 있기 때문에 N은 여파기의 통과대역율이다.

그리고 처리프레임 마다 성문펄스의 모양과 포먼트 주파수가 변화하기 때문에 여파기의 차단주파수를 고정할 경우 포먼트의 영향을 제대로 제거할 수 없게 된다. 따라서 G-peak검출시 포먼트의 영향을 적응적으로 제거하기 위해 식 (3-1)의 저역통과여파기의 차단주파수 f_T를 처리프레임 마다 가변시켜야 한다.

본 논문에서는 여파기의 차단주파수를 G-peak의 특성을 이용하여 처리 프레임에 따라 가변적으로 구하여 적용하였다. 즉, G-peak는 성문펄스의 파형과 제 1포먼트의 포락선이 서로 편법투선된 신호의 첫피크이기 때문에 G-peak에서의 영교차간격이 인근 피크들에서의 영교차간격 보다 길다는 특성을 갖는다. 그러므로 음성신호 S(.)에 대해서 영교차점을 검출하여 각 영교차점간의 간격 ZCI(.)를 다음 식과 같이 계산한 뒤, 최대인 ZCI(.)를 여파기의 통과대역율 N으로 결정한다 :

$$ZCI(i) = Z_c(i+1) - Z_c(i) \quad \cdots\cdots \quad (3-2)$$

여기서 $Z_c(i)$는 i번째 영교차점을 나타내고, $Z_c(i+1)$은 i+1번째 영교차점을 나타낸다.

여파기를 통과한 신호 S'(.)은 그림 3-1(b)에 나타낸 것처럼, 포먼트 성분을 갖는 인근 피크들은 감소되는 반면, 성문파의 특성을 나타내는 G-peak는 부각되어 나타나게 된다.

이때 S'(.)의 진폭은 $|\max(S'(.))| > |\min(S'(.))|$ 가 된다. 여기서 G-peak의 봉우리는 GND를 기준으로 max쪽으로 크게 부각되어 나타나게 되는 반면, 포먼트의 영향을 받는 인근 피크들의 봉우리는 GND를 기준으로 그 근방과 min쪽으로 치우쳐 감소되어 나타나게 된다. 따라서 여파된 신호 S'(.)에서 G-peak만을 검출하기 위해 다음과 같이 문턱값 T를 결정한다 :

$$T = |\min(S'(.))| \quad \cdots\cdots\cdots\cdots \quad (3-3)$$

결정된 문턱값에 의해 검출된 봉우리가 G-peak를 나타내며, 이를 그림 3-1(c)에 나타내었다. 그리고 G-peak 검출에 의해 결정된 피치시점을 그림 3-1(d)에 나타내었다.

Ⅳ. 실험 및 결과

이상의 것을 컴퓨터 시뮬레이션하기 위해 마이크가 장치된 12-bit A/D변환기를 IBM-PC/486에 인터페이스 시키고, 아래의 발성을 8KHz의 샘플링 주파수로 양자화하여 저장한 다음 시뮬레이션에 대한 시료로 사용하였다.

발성1: "인수네 교마는 천재소년을 좋아한다."
발성2: "예수님께서 천지창조의 교훈을 말씀하셨다."
발성3: "숭실대 정보통신공학과 음성처리 연구실이다."
발성4: "공일이삼사오육칠팔구."

각 음성시료에 대해 한 프레임의 길이를 256샘플로하여 128샘플 단위로 오버랩하여 처리하였다.

음성신호 S(.)에 대해 직접 피치시점을 검출할 경우 상위포먼트들의 영향으로 검출에러가 많이 발생하며, 또한 결정논리가 복잡하게 된다. 따라서 상위 포먼트의 영향을 제거하면서 G-peak를 강조하기 위해 음성신호를 저역통과 여파기에 통과시킨 뒤에 피치시점을 검출하였다. 이때 G-peak의 특성을 이용하여 처리프레임마다 여파기의 통과대역율을 가변적으로 하였다. 즉, 음성신호에서 영교차점간의 간격을 측정하여 최대 영교차간격을 여파기의 통과대역율로 사용하였다. 그리고 여파기를 통과한 신호를 그림 3-1(b)에 나타내었다.

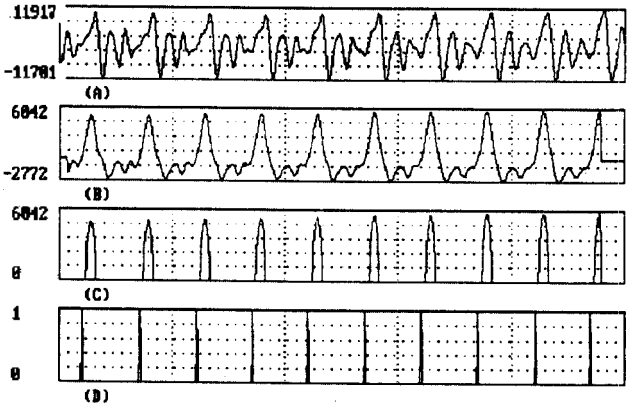

그림 3-1. G-peak의 특성을 이용한 피치시점검출 과정
(a) 음성파형, (b) 가변 LPF를 통과한 파형,
(c) 검출된 G-peak, (d) 검출된 피치시점

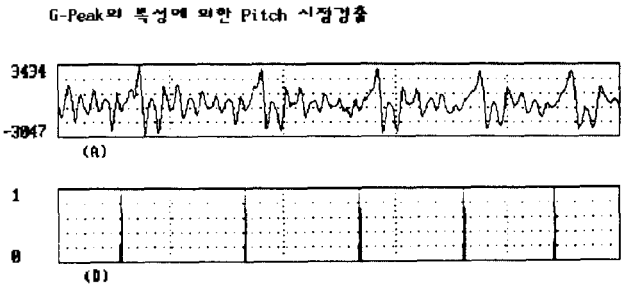

그림 4-1. 제안한 방법으로 검출한 피치시점
(a) 음성신호, (b) 검출된 피치시점

여파기를 통과한 신호 S'(.)는 그림 3-1(b)에서 처럼 G-peak가 부각되어 나타나며, 인근 피크들은 GND를 기준으로 감소되어 나타난다. 따라서 S'(.)의 최소값을 구해 $|\min(S'(.))|$을 문턱값으로 사용하여 G-peak만을 검출하였다. 검출된 G-peak를 그림 3-1(c)에 나타내었다. 그리고 검출된 G-peak의 영교차점을 피치시점으로 검출하였으며, 검출된 피치시점을 그림 4-1(b)에 나타내었다.

V. 결 론

음성신호처리 분야에서 피치시점을 정확히 검출하는 것은 아주 중요하다. 만약 피치시점을 정확히 검출할 수 있다면, 음성분석시 피치에 동기시켜 분석할 수 있고, 인식시에는 성문의 영향이 제거된 정확한 성도 파라미터를 얻을 수 있으므로 인식의 정확도를 높일 수 있다. 또한, 합성시에는 여기원의 위상특성을 파악할 수 있으므로 개성이 강조된 합성음을 얻을 수 있다.

본 논문에서는 음성신호의 발성모델에 근거하여 G-peak를 검출한 다음, 이를 이용하여 피치시점을 검출하는 새로운 피치시점 검출 알고리듬을 제안하였다.

제안한 방법은 시간영역에서 직접 피치시점을 검출하므로 처리 알고리듬이 간단하고, 음성신호의 위상성분을 그대로 유지하면서 정확히 피치시점을 검출 할 수 있었다.

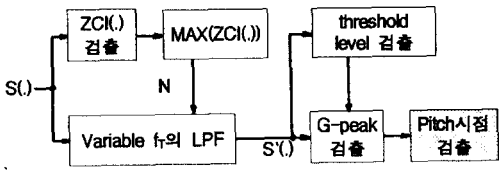

그림 4-2. 처리과정에 대한 블럭도

참고 문헌

[1] L.R.Rabiner and R.W.Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[2] J.D.Markel and A.H.Gray, jr., *Liner Prediction of Speech Signals*, Springer-Verlag, 1976.

[3] DALE E.VEENEMAN, SPENCER L.BEMENT, "Automatic Glottal Inverse Filtering from Speech and Electroglottographic Signals", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol.ASSP-33, No.2, pp.369-377, APRIL 1985.

[4] YAN MING CHENG, DOUGLAS O'SHAUGHNESSY, "Automatic and Reliable Estimation of Glottal Closure Instant and Period," IEEE TRANSACTIONS ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, Vol.37, No.12, December 1989.

[5] Hans Werner Strube, "Determination of the instant of glottal closure from the speech wave", J.Acoust.Soc.Am., Vol.56, No.5, pp.1625-1629, November 1974.

[6] A.E.Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels", J.Acoust.Soc.Am., Vol.49, pp.583-590, 1971.

[7] L.R.Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection", IEEE Trans on Acoustics, Speech, and Signal Proc., Vol.ASSP-26, No.1, pp.24-33, February 1977.

[8] 배명진, 임재열, 안수길, "음성발생 모델로 부터의 G-peak를 이용한 음성에너지 추출에 관한 연구", 전자공학회논문지, Vol.24, No.3, pp.381-386, 1987.

[9] D.Y.Wong, J.D.Markel, and A.H.Gray, Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-27, PP.350-355, Aug. 1979.

[10] Myungjin BAE and Souguil ANN, "A study on the fundamental frequency extraction of speech signals using second order rundown method", Seoul National university, MA Paper, Jan. 1983.

[11] G.Fant, *Acoustic Theory of Speech Production*. Gravenhage, The Netherlands:Mouton, 1960.

[12] 이해군, 배명진, 임윤천, "G-peak 검출에 의한 음성신호의 피치시점검출", 제6회 신호처리합동학술대회논문집, Vol.6, No.1, pp.58-61, 1993.