고조파 분석에 의한 음성 신호의 피치 검출

°김 기 희　　최 정 아　　배 명 진　　안 수 길
서울대학교 공과대학 전자공학과

Pitch Extraction of Speech Signals by the Harmonics analysis

Keehee KIM　Jungah CHOI　Myungjin BAE　Souguil ANN
Seoul National University

The harmonics of the fundamental frequency in speech signal make a minute line spectrum in frequency domain. In this paper, we propose a new algorithm to detect a pitch interval in voiced sound based on the fact that the number of harmonics can represent the period of the pitch in the time domain.

I.서 론

음성 신호처리 분야에서 음성의 특징을 추출하기 위한 분석과정은 응도에 따라 다르게 선택되고 있다. 음성의 발성 모델을 필터적인 관점으로 처리하려는 source 코딩법은 음원 부분을 필터의 excitation으로, 성도 부분을 필터의 응답 특성으로 취급하고 있다. 음원 부분은 무성음, 유성음에 따라 random noise나, 임펄스 열을 사용한다. 특히 유성음에서 임펄스 열은 피치의 간격으로 공급되어야 하며, 피치(pitch)는 화자의 개성이나 음의 자연감을 나타내는 중요한 패러미터가 된다.

지금까지 연구된 방법은 크게 세가지로 나눌 수 있는데 시간 영역의 피치 검출기, 주파수 영역의 피치 검출기 및, 두 영역 혼합 방식이 있다.

시간 영역에서의 검출 방식은 우선 음성 파형에서 주기 가능한 성분을 검출하고 나서, 결정 논리회로에 의해 실제의 피치를 검출해 내는것이다. 이것은 시간 영역의 음성 파형에서 직접 추출하기 때문에 계산 시간이 단축된다는 장점이 있지만 음소의 천이 구간(transition interval)에서는 파형의 급한 변동 때문에 피치의 helving이나 doubling이 발생할 수 있다. 따라서 이러한 문제를 해결하려면 피치를 결정하는 논리가 복잡해져서 결국은 별도의 필터링이 필요하게 된다[2].

주파수 영역에서의 검출 방식은 음성의 파형 성분이 음원과 필터 응답이 서로 convolution되어 발생된다는 사실을 이용하고 있다. 시간 영역에서의 convolution은 주파수 영역의 곱셈으로 나타나고, logarithm을 적용한 후에는 덧셈으로 처리될 수 있기 때문에 기본주파수의 성분을 분류해 낼 수 있게된다[2].

주파수 영역에서의 검출 방식은 프레임 단위로 음성신호를 잘라서 처리하기 때문에 averaging 효과에의해 그 프레임 안의 미세한 피치들의 변화는 나타낼 수 없다. 그리고 다른 영역으로의 변환 과정이 적용되기 때문에 변환 과정에서 고 정도의 연산이 필요해지며, 계산량이 많아지게 된다. 그럼에도 최근 신호 처리용 IC들의 등장에 따라 곱셈의 계산 시간이 덧셈의 계산 시간 같게 적용되기 때문에 피치 검출기의 복잡성 보다는 정확한 검출 방법의 제안이 요구되고 있다.

피치를 정확히 추출하기 어려운 구간으로는

1. 음소의 천이 구간

2. 유성음의 앞뒤에 따라오는 낮은 에너지의 유성 자음구간

3. 유성음의 앞뒤에 연결되는 비음 구간

이 있으며, 남녀 노소에 따른 피치의 변화를 잘 추종할 수 없다는 어려운 문제점들이 있다. 이러한 문제를 해결하려면 주파수 영역과 시간영역에서의 분석법을 함께 적용하는 hybrid법이 적용되어야 함을 암시한다. 그렇지만 hybrid법은 별로 제안되지 않고 있다.

II. 스펙트럼의 분석

음성 신호의 발생 과정을 음원에 따라 나누어 선형 시스템으로 모델링해보면 그림 2-1 과 같다[3]. 이 모델에 의하면 유성음은 성대의 진동에 따른 준주기적 성질을 펄스열로 하고 이 음원이 필터로 모델링된 성도를 통해서 나오는 것으로 나타낼 수 있음을 알 수 있다.

이것을 수학적으로 표현해 보면 다음과 같다.

$S(t) = W(t) * i(t)$

dt : sampling 간격

i(t) : 음원

W(t) : 필터의 impulse 응답

S(t) : 음성 신호

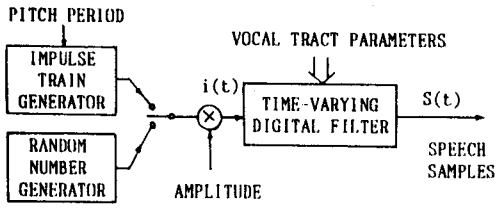

그림 2-1 음성 신호 발생 모델

유성음의 음원은 준주기적 성질을 갖는다. 이를 피치(pitch)라 하는데 유성음 신호 S(t) 의 주파수가 $f < 1/2dt$ 로 제한되어 있다면 유한한 길이의 Fourier Series 로 나타낼수 있다.

$$S(t) = \sum_{k=-K}^{K} C_k e^{jk\frac{2\pi}{T_0}t}$$

$$C_k = \frac{1}{T_0}\int_0^{T_0} S(t)\, e^{-j\frac{2\pi}{T_0}kt}\, dt$$

$$Z(f) = \int_{-\infty}^{\infty} S(t)\, e^{-j2\pi ft}\, dt$$

$$= \int_{-\infty}^{\infty}\left(\sum_{k=-K}^{K} C_k e^{jk\frac{2\pi}{T_0}t}\right) dt$$

$$= \sum_{k=-K}^{K} C_k \int_{-\infty}^{\infty} e^{-j2\pi(-\frac{k}{T_0}+f)t}\, dt$$

$$= \sum_{k=-K}^{K} C_k\, \delta(f-\frac{k}{T_0})$$

Z(f) 에서 보면, 주파수 영역에서도 역시 주기성이 나타남을 알 수 있다.

III. 피치 의 검출

주파수 영역으로 변환된 신호의 고조파 의 갯수를 세기 위해서는 임의의 threshold 를 선택한 후 이를 crossing 하는 갯수를 세면 된다. 그러나 주파수 영역에서의 신호는 threshold 의 선택이 용이한 파형은 아니므로 파형의 영향이 적은 방법을 선택해야 한다. 이러한 threshold 에 영향을 줄 수 있는 문제점과 해결 방법은 다음과 같다.

시간 영역에서 음성 신호를 한 프레임을 취해서 주파수 영역으로 변환하면 각 주파수의 절대치가 같지 않으므로 envelope 에 변화가 나타나게 된다. 이러한 envelope 의 변화가 크면 threshold 의 선택에 어려움을 주게 된다. 또 주파수 영역의 파형에는 미세한 변화가 나타나게 된다. 만약 선택된 threshold 부근에 이러한 미세 변화가 나타날 때 이 threshold 를 crossing 하는 갯수를 센다면 그릇된 결과를 얻을 수 있다.

그러므로 이러한 문제점들을 제거하기 위해서 threshold 는 주파수 영역 파형의 average 를 잡고,

이것을 crossing 하는 갯수를 센다. 높은 주파수 영역에서는 aliasing 의 영향 때문에 고조파의 봉우리가 불분명하게 나타난다. 그러므로 이를 피하기 위해서 먼저 avreage 를 crossing 하는 갯수가 10 개 되는 점을 구한다. 이점의 위치를 i 라 하고, 한 프레임 내에서의 피치 간격이 일정하다고 가정하면,

$$\text{피치 간격} = \frac{256}{i} * 10$$

에서 피치 간격을 구할 수 있다.

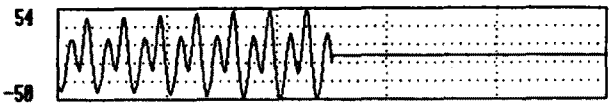

그림 2-a 음성 신호 "오"의 파형

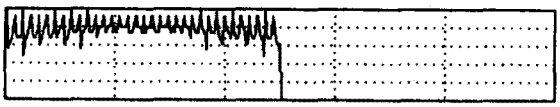

그림 2-b 음성 신호 "오"의 면적에 대한 라인 스펙트럼

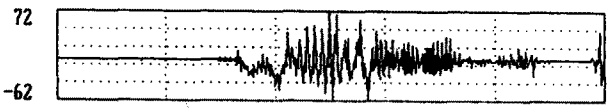

그림 3-a 음성 신호 "MAY-I-HELP-"의 파형

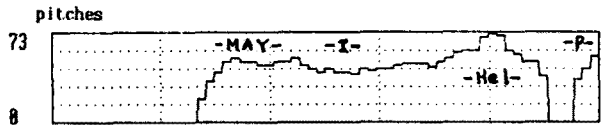

그림 3-b 음성 신호 "MAY-I-HELP-"의 피치 변화

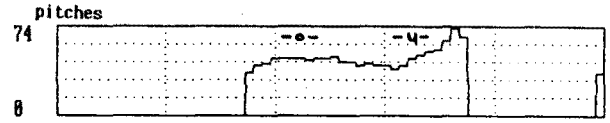

그림 4 음성 신호 "오"에 대한 피치 변화

IV. 실험및 결과

음성신호는 유성음과 무성음으로 구분되고, 피치는 유성음 구간에서 만 존재한다. 따라서 피치를 검출하기 전에 먼저 유음과 무음 구간을 구분한다[12]. 그리고 나서 유음구간에 대해 유성음 구간과 무성음 구간을 분류한다[10]. 유성음 구간이 찾아지면 프레임 단위로 피치를 찾는 그림1의 알고리즘을 적용한다. 시뮬레이션에 사용한 데이타는 음성 신호를 8KHz로 표본화하였으며 한 프레임의 길이를 256-samples(=32 msec)로 하여, 한 프레임 마다 겹치게 분석하였다.

유성음의 한 프레임이 결정되면 FFT를 적용하기 전에 실수값 버퍼X(.)에는 음성 파형의 각 봉우리가 끝나는 곳에다 그 봉우리의 면적값을 넣고, 허수값 버퍼y(.)에는 영을 넣는다. 이러한 면적들을 대신 사용하면 음성 신호의 피치 성분이 더 강조될 수 있다[9]. 이러한 값에 대해 256-포인터 FFT를 수행하고 나서는 위상특성 보다는 그 고조파 성분을 조사하기 위해 에너지를 계산해야 한다. 에너지 스펙트럼은 Logaritha 형태로 많이 이용하지만 우리의 관심사는 고조파의 갯수를 파악하는데 있기 때문에 자승된 에너지를 그대로 사용하였다 (플로우-차트에서X(.)*X(.)+Y(.)*Y(.)).

시간영역에서 한 피치가 차지하는 구간에 대한 sample의 개수는 스펙트럼에서 나타나는 고조파들의 갯수와 같다. 그럼지만 고조파의 개수를 파악하려면 높은 주파수 쪽으로 갈수록 aliasing 현상에 의해 파악하기가 어렵게된다. 따라서 고조파들이 두렷이 나타나는 낮은 주파수 영역에서의 값으로 높은 주파수 영역을 대치할 수 있다. 영-주파수에서 부터 10-번째의 고조파가 나타나는 주파수값 i를 알으면 피치 주기를 sample의 개수로 나타낼 수 있게된다. 즉,

$$\text{pitch} = (256/i) * 10 \quad [\text{samples}]$$

고조파들은 그림2에서 처럼 주파수 구간에서 정현파 형태를 이루고 있다. 고조파의 개수를

카운트하려면 정현파 모양의 중간값을 기준으로 하여 이것이 -에서 +로 교차하는 개수를 파악하면 된다. 기준치를 나타내는 중간값 Thr은 스펙트럼 상에서 기본주파수가 제일 높은 경우를 생각하여 12개의 구간(=400Hz/(8000Hz/256points)) 의 평균값을 사용하였다.

이상의 과정을 통하여 음성신호 "May I help you?"에 대한 피치 검출 결과를 그림3에 나타내었다. 그림3을 발음한 화자는 25세의 남성이었다. 그림4는 23세의 여성에 의해 발음된 고립 숫자음 "오"에 대한 결과이다. 그림3의 경우는 발음 중간에 파열음이 들어 있기 때문에 시간 영역에서는 음소의 천이 구간이 나타나는 경우이다. 그렇지만 피치의 halving이나 doubling없이 피치가 잘 찾아지고 있음을 알 수있다. 특히 기존의 방법에서는 피치의 변화 그래프가 연속적으로 찾아지지 않았기 때문에 smoothing을 필요로 했지만 그것을 적용할 필요가 없어진다. 이것은 주파수 영역으로의 변환 과정에서 이미 smoothing이 되었기 때문이다.

V. 결 론

유성음은 음원을 이루는 펄스열이 성도의 특성을 갖는 필터를 통과해서 얻어진다고 할 수 있다. 이 때문에 시간영역에서의 유성음의 파형은 피치 주기의 특징이 두드러지게 된다. 주파수 영역에서도 기본주파수의 고조파 성분들이 새새하게 나타나서 라인 스펙트럼을 형성하게 된다. 본 논문에서는 고조파 성분의 갯수가 시간영역의 피치 주기를 나타낸다는 점에 착안하여 이 갯수를 검출함으로써 유성음의 피치를 구하는 새로운 알고리즘을 제안하였다.

한 프레임은 보통 피치주기의 두배 이상이 되도록 선택되는데 그 프레임 안에서 피치의 주기가 변화하고 있다면 고조파 성분들 사이에 미세한 구조로 나타나게

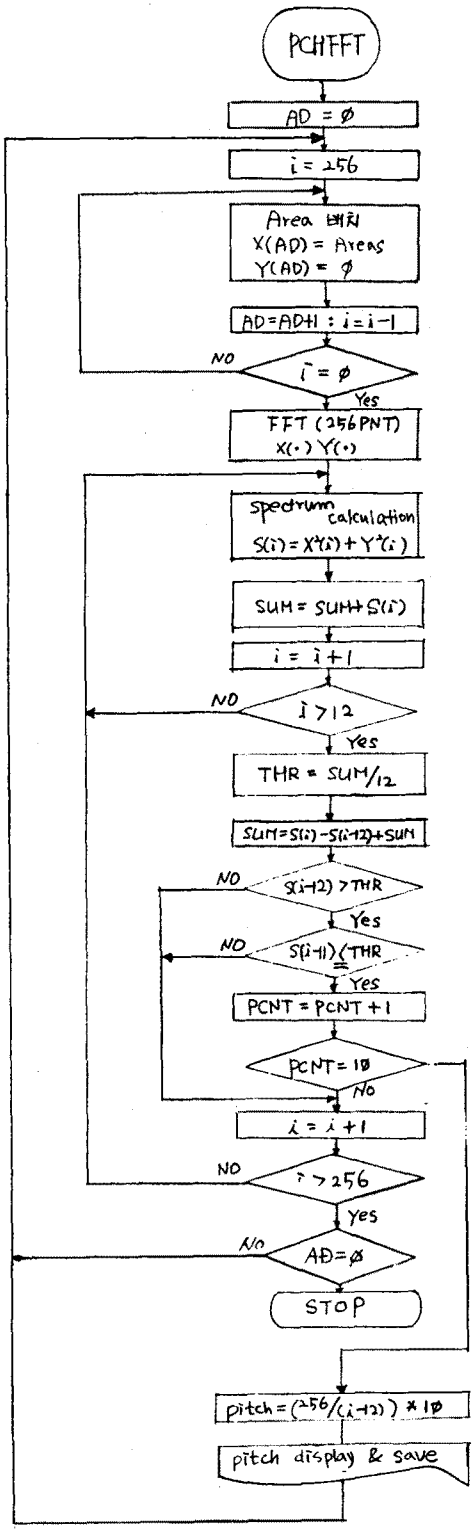

그림 1 피치 검출에 대한 전체적인 플로우차트

된다. 이러한 미세구조에 의해 고조파의 갯수를 잘못파악할 수도 있게된다. 따라서 우리는 시간영역에서 파형 그 자체를 이용하는것 보다는 파형을 구성하고 있는 봉우리들의 면적을 주파수 영역으로 변환시킴으로서 주기성을 강조시켰다. 또한 낮은 주파수쪽의 고조파 성분이 높은 주파수쪽의 것 보다 분명히 나타나기 때문에 낮은 주파수쪽의 고조파 성분을 대표적으로 사용하는 방법을 제시하였다.

제안된 피치 검출방법은 검출시에 문제가 되는 halving이나 doubling을 제거하기위한 까다로운 길정 알고리즘이나 smoothing 알고리즘이 요구되지 않는다고 할 수 있다. 특히 화자에 대해서는 남여 노소에 무관한 피치 검출 방법이 될수 있다.

VI. REFERENCES

[1] J.D.Markel and A.H.Gray, Linear Prediction of Speech, Springer Verlag, Berlin, 1976.

[2] L.R.Rabiner and R.W.Schafer, Digital Processing of Speech Signals, Prentice-Hall, Elglewood Cliffs, New Jers- ey, 1978.

[3] B. Gold and L.R. Rabiner, "Parallel Processing Technique for estimating Pitch Periods of Speech in the Time Domain", J Acoust. Soc. Am., Vol.46, No.2, PP. 442-448, August 1969

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

[5] J.D.Markel, "The SIFT Algorithm for Fundamental Frequency Estimation" IEEE Trans. on Audio and Electraceustics, Vol. Au-20, No 5, pp367-377, December, 1972.

[6] M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, and H.J. Manley, Average Magnitude Difference Function Pitch Extractor", IEEE Trans, Acoust. Speech and Signal Proc., Vol. ASSP-22, PP.535-562, Oct. 1974

[7] N.C. Geckinli, D. Yavuz,"Algorithm for Pitch Extraction using Zero Crossing Interval Sequence", IEEE Trans. Acoust. Speech and Signal Proc., Vol.ASSP-25, No.6, PP. Dec.1977

[8] Myungjin BAE,"A Study on the Fundamental Frequency Extraction of Speech Signals using Second Order Rundown Method",Seoul National University, MA Paper,Jan. 1983

[9] Myungjin BAE and Souguil ANN, "The High Speed Pitch Extraction of Speech Signals using the Area Comparison Method", KIEE, Vol. 22, No. 2, PP. 101-105, Feb. 1985.

[10] Myungjin BAE and Souguil ANN, "The Voiced-Unvoiced-Silence Classification by Emphasized Spectrum of Speech Signals",J. of Acoust. the Acou. Soc. of Korea, Vol. 4, No.5,pp9-15,June 1985

[11] Myungjin BAE and Souguil ANN,"Inverse Rate Type Filtering for the Pitch Extraction",J. of the Acou. Soc. of Korea, Vol.5, No.3, PP.5-12 , Sept. 1986

[12] Myungjin BAE, Souhwan JUNG, and Souguil ANN "End-Point Detection of Speech Signals by Bit Crossing Rate", to be published, 1987