웨이브렛 변환을 이용한 피치검출

손 영호, 석 종원, 배 건성

경북대학교 전자·전기 공학부

Pitch Detection Using Wavelet Transform

Young Ho Son, Jong Won Seok, Keun Sung Bae

School of Electronic and Electrical Engineering, Kyungpook National University

요 약

음성신호는 성대를 통과한 공기 흐름의 성질에 따라 유성음과 무성음으로 구분될 수 있다. 그 중 유성음에서는 성대의 규칙적인 진동이 존재하게 되는데 이때 성대가 닫히는 순간을 GCI(Glottal Closure Instant)라 하며 성대 진동의 기본주기를 피치라고 한다. 이러한 피치는 음성합성, 음성인식, 피치동기 음성신호의 분석 등에 중요한 파라미터 중의 하나이다. 본 논문에서는 가우시안 함수의 일차 미분형태를 갖는 웨이브렛 함수를 사용할 경우 신호의 급격한 변화부분을 검출할 수 있다는 성질을 이용하여 음성신호의 GCI를 찾아내고 이를 이용하여 피치를 검출하였다.

Ⅰ. 서 론

음성신호는 인간의 발성기관에서 2개의 얇은 막으로 구성된 성대(vocal folds)의 자발적인 운동에 의해 발생되는 공기의 흐름이 성도(vocal tract)를 지나면서 변조되어 공기압의 파동형태로 나타나는 것이다. 이러한 음성신호는 성대를 통과한 공기흐름의 성질에 따라 크게 유성음(voiced sound)과 무성음(unvoiced sound)으로 나눌 수 있다. 모음과 같은 유성음을 발성할 경우 허파에서 방출되는 공기는 닫혀진 성대에 의해 공기압이 점차 증가하여 성대가 떨어지기 시작하면서 좁은 공기 통로를 형성하게 되는데 이를 성문(glottis)이라고 하며, 성문을 통과하는 공기는 서로 다른 두 힘의 상호작용에 의해 성대가 규칙적으로 진동을 하도록 만든다. 이처럼 유성음에는 성문이 닫혀져 있는 부분과 열려져 있는부분이 있게 되는데 성문이 닫히는 순간을 GCI 또는 epoch이라 한다. 유성음을 발성할 때의 성대의 단위시간당 진동횟수 즉, 기본주파수 또는 반복되는 진동운동의 기본주기 T_0를 피치라고 하는데, 음성합성, 음성인식, 화자인식, 화자검증, 피치동기 음성신호의 분석 및 합성 등 음성신호처리 분야에 있어서 매우 중요한 파라미터 중의 하나이다.[1]

피치를 검출하는 방법에는 크게 event detection 방법과 non-event detection 방법이 있는데 이 때 event란 epoch을 의미한다. non-event detection 방법이 자기상관함수 또는 AMDF 등의 방법으로 평균적인 피치를 검출하는 방법인데 반해 event detection 방법은 먼저 epoch을 검출한 후 연속적인 epoch들 사이의 시간적 간격을 측정함으로써 피치를 검출하는 방법을 말한다. 최근에 음성신호를 웨이브렛 변환한 신호에서 local maxima가 실제 음성에서 급격한 변화를 나타내는 GCI에 해당된다는 결과가 발표되었으며 이를 이용한 피치 검출방법은 피치주기의 non-stationary한 변화부분에 대해서 뿐만 아니라 다양한 화자들 사이의 변화에 대하여서도 잘 대처할 수 있는 방법으로 알려져 있다.[2-4] 본 논문에서는 피치를 검출하는 방법으로서 Mallat이 제안한 quadratic spline 웨이브렛 함수를 이용하여 epoch을 검출하고 그 결과를 이용하여 음성의 피치를 구하였다. 이때 검출된 epoch은 epoch 검출에 있어서 기준신호로 적합한 것으로 알려진 EGG(Electroglottograph) 신호와 비교하여[6-8] 타당성을 조사하였으며, 그리고 이에 따른 피치검출 알고리즘을 연구하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 웨이브렛 변환에 대하여 설명하며 3장에서는 quadratic spline 웨이브렛 함수를 이용한 피치검출 방법 대하여 설명한다. 그리고 4장에서는 3장에서의 방법으로 실험한 결과를 미분된 EGG(DEGG) 신호와 비교 설명하며 5장에서 결론을 맺는다.

Ⅱ. 웨이브렛 변환

웨이브렛 변환은 응용수학에서 처음 소개된 후 최근 컴퓨터비전 분야 등에서 연구되어 온 다중 해상도 표현과 연관성이 있음이 밝혀졌으며 이산 웨이브렛 변환 이론은 이산 신호의 subband 분할 방법과도 연관성이 존재한다. 그 중에서도 계수 구현을 더욱 용이하게 하는 dyadic 웨이브렛 변환(DyWT: Dyadic Wavelet Transform)은 식 (1)과 같고 이때 웨이브렛 함수는 식 (2)와 같이 정의된다.

$$W_2^{d,j} = \frac{1}{\sqrt{2^j}} \int f(t)\, \varphi^*\left(\frac{t}{2^j} - kT\right) dt \qquad (1)$$

$$\varphi_{j,k}(t) = 2^{\frac{-j}{2}} \varphi(2^{-j}t - kT) \quad (2)$$

식 (1)의 웨이브렛 변환 $W_2^{d/b}$에서 d는 이산변환을 가리키며 j는 scale을 나타낸다.

웨이브렛 변환을 이용하여 신호를 분석하는 과정은 그림 1에서 보는 것과 같은 tree 형태의 필터뱅크로 생각할 수 있다. 여기서 H_0은 저역통과 필터이고, H_1은 고역통과 필터이다. 입력신호가 저역통과 필터와 고역통과 필터를 거치게 되면 한 번의 웨이브렛 변환이 이루어지며, 저역필터를 통과한 신호에 대해 이러한 과정을 반복적으로 수행하여 웨이브렛 변환된 신호를 얻을 수 있다. 신호처리 관점에서 볼 때 DyWT는 constant-Q, octave band, 밴드패스 필터들의 뱅크 출력과 동일하다.

본 연구에서 사용한 Mallat에 의해 제안된 quadratic spline 웨이브렛 함수는 식 (2)에서 정의된 웨이브렛 함수를 smoothing 함수 $\theta(t)$의 일차 미분으로 할 때 DyWT의 local maxima는 신호에 있어서 급격한 변화부분을 가리키고 local minima는 느린 변화를 나타낸다는 것을 입증하였다. 이 때 smoothing 함수란 어떤 함수의 Fourier 변환에서 저주파에 에너지가 몰려있는 함수를 말한다. 더욱이 Mallat은 시간 $t=t_0$에서 실제 신호가 갑작스런 변화를 하게될 때 $t=t_0$에서 연속적인 scale에 걸쳐서 DyWT한 값들은 local maxima를 가지게 된다는 것도 입증하였다.[5]

본 연구에서 사용한 웨이브렛은 Mallat의 quadratic spline 웨이브렛으로서 사용된 필터의 계수는 표 1과 같다.

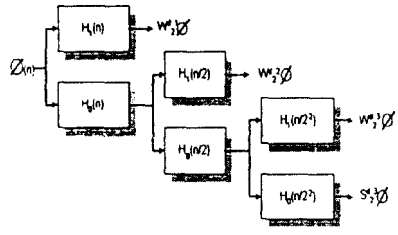
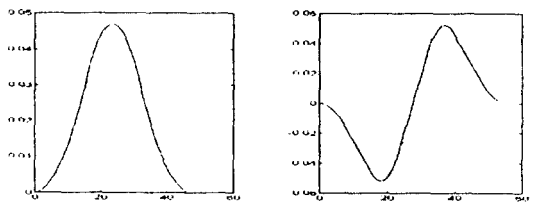


그림 1. 웨이브렛 분해 필터뱅크

표 1. quadratic spline의 필터 계수

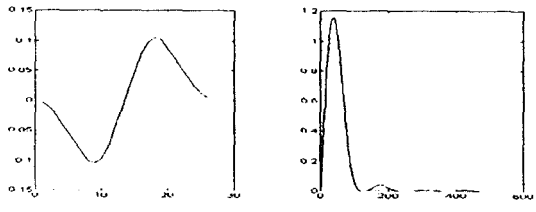
n	H_0	H_1
-1	0.125	
0	0.375	-2.0
1	0.375	2.0
2	0.125	

그림 2는 본 논문에서 사용된 smoothing 함수와 웨이브렛 함수의 관계를 보여주고 있으며 그림 3에서는 각각 스케일 3과 4에서의 웨이브렛 함수와 주파수 영역에서의 스펙트럼 특성을 보여주고 있다. 그림 2에서는 (a)의 smoothing 함수를 미분한 함수가 (b)의 웨이브렛 함수와 위상만이 반전된 동일한 함수임을 확인할 수 있다. 그리므로 epoch의 검출은 음의 값에서의 local maxima값이 검출되게 된다. 본 논문에서의 local maxima와 global maxima는 모두 음의 방향의 값들이다. 그림 3과 4에서의 FFT 스펙트럼에서는 스케일 4에서의 주파수 영역이 스케일 3에 비해 저주파 영역쪽으로 절반 정도임을 볼 수 있다. 이 사실은 실제 음성의 웨이브렛 변환에서 스케일 3에서의 신호보다 스케일 4에서 얻어진 신호가 피치주기 성분을 더 반영할 수 있음을 의미한다.



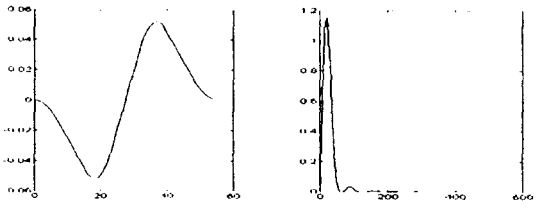
(a) (b)

그림 2. 스케일 4에서 (a) smoothing 함수, (b) 웨이브렛 함수



(a) (b)

그림 3. 스케일 3에서 (a) 웨이브렛 함수, (b) FFT 파형



(a) (b)

그림 4. 스케일 4에서 (a) 웨이브렛 함수, (b) FFT 파형

III. Quadratic Spline을 이용한 피치 검출

Quadratic spline을 이용한 웨이브렛 변환에서는 일반적인 subband 분할 방식과는 달리 필터링된 신호의 크기를 그대로 유지하면서 다음 스케일 신호를 얻기 위해 필터링된 신호를 decimation하는 대신 필터계수들 사이에 0을 삽입하여 만든 웨이브렛 함수의 신호를 컨볼루션한다.[5] 앞에서 언급한 quadratic spline 웨이브렛 함수의 성질을 이용한 피치 검출 알고리즘은 아래와 같다.

STEP 1: 먼저 전체 음성을 스케일 3, 4, 5에서의 웨이브

렛 함수를 가지고서 각각 웨이브렛 변환한다. 이 때 웨이브렛 함수와 음성신호와의 convolution을 하게되는데, 각 스케일에서의 필터링으로 인한 delay(각 스케일에서의 웨이브렛 길이,W_i의 1/2)를 고려해주었다.

STEP 2: Step 1에서의 웨이브렛 변환한 신호를 분석하기 위하여 본 실험에서는 150 샘플씩 윈도윙하여 분석하도록 하였다. 먼저 해당 구간에 대하여 스케일 3에서의 global maxima를 구하여서 유성음과 무성음을 구분하는 threshold, T_0와 비교를 하여 작을 경우는 묵음이나 무성음 구간으로 간주하고 다음 구간으로 넘어간다. 이 때 threshold를 잡는 방법은 유성음 구간에서의 웨이브렛 변환한 값이 무성음이나 묵음 구간에 비해서 훨씬 크다는 것을 고려하여 적응적으로 잡도록 하였다.

STEP 3: Step 2에서 구간의 global maxima 값들이 T_0보다 클 경우에는 스케일 3과 4에서의 local maxima를 찾아서는 스케일 3에서는 T_3×global $maxima_3$, 스케일 4에서는 T_4×global $maxima_4$를 넘어서는 local maxima값들을 찾아서 그 값들의 위치가 서로 같은지를 비교하고 스케일 5에서의 값도 epoch의 검출에 이용하였다. 이 때 위치의 허긋남이 10 sample 이내고 스케일 5에서의 크기 조건을 만족하게 될 경우 일치하는 것으로 간주하고 값의 위치를 검출된 epoch으로 간주한다.

STEP 4: 한 프레임의 분석이 끝나면 다음 프레임으로 이동한다. 이 때 각 프레임은 10샘플씩 겹치도록 하였다. 이것은 윈도우의 끝 부분에서 epoch에 대한 정보를 잃어버릴 경우의 영향을 보상해주기 위한 것이다. 그리고서는 다시 Step 1로 돌아가서 위의 단계를 반복한다.

STEP 5: Step 1-4에서 전 음성구간에 대해서 검출한 epoch을 가지고서 연속된 epoch들 사이의 시간적 간격을 측정함으로써 피치를 구한다.

본 연구에서는 실험적으로 유성음과 무성음을 구분하는 threshold, T_0는 알고리즘 자체에서 결정하도록 하였으며 T_3, T_4는 각각 0.45와 0.5로 두었으며 epoch의 검출에서는 스케일 4에서의 신호를 가준으로 실험을 하였다.

Ⅳ. 실험 및 검토

본 연구에서는 남자 3명, 여자 3명의 화자에 의하여 10 kHz로 샘플링된 음성을 대상으로 하여 실험을 하였다. 사용된 문장은 아래와 같다.

Sentence A : "We saw the ten pink fish."
Sentence B : "We were away a year ago."
Sentence C : "Early one morning a man and a woman ambled along a one mile lane."
Sentence D : "Should we chase those cowboys? "

그림 5에서 7까지는 실제 음성신호, DEGG 신호, 스케일 3과 스케일 4, 5에 해당하는 웨이브렛 변환된 신호를 비교하여 나타냄으로써 웨이브렛 변환을 이용하여 검출한 epoch이 실제의 epoch과 일치함을 보여주고 있다. 그림 5에서는 voice onset 구간에 대하여 각 신호들을 비교하고 있으며, 그림 6에서는 안정된 유성음 구간에 대하여, 그림 7에서는 voice offset 구간에 대하여 각각 나타내고 있다. 그림에서 음성신호와 DEGG 신호는 시간축에 대하여 동기되어져 있다. 각 그림에서 스케일 5에서 변환된 신호는 스케일 3, 4에서 변환된 신호에 비해 많이 smoothing 되었으며 smoothing으로 인하여 local maxima의 위치의 변화량이 스케일 3, 4에서 위치의 변화량보다 크게 나타나는 것을 확인할 수 있다. 그러나 실제의 epoch과 일치하는 부분에서는 반드시 음의 값을 갖게됨을 확인 할 수 있다. 또 각 스케일에서의 local maxima값의 변화량들은 음성의 안정된 유성음 구간에서보다 voice onset이나 voice offset 구간에서 상대적으로 더 크게 나타나고 있다.

그림 8은 검출한 epoch을 이용하여 구한 음성의 피치 contour를 나타낸 것이며, 그림 9는 웨이브렛을 이용한 피치 검출이 피치주기의 non-stationary한 경우에서도 잘 대처할 수 있다는 예를 보여주고 있다.

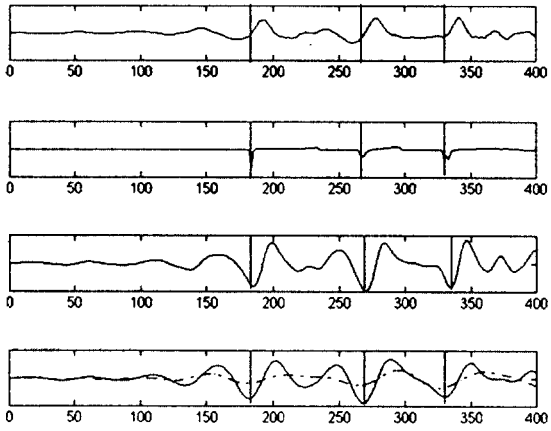


그림 5. voice onset 구간의 예
(speech/DEGG/scale3/scale4(-), scale5(-.))

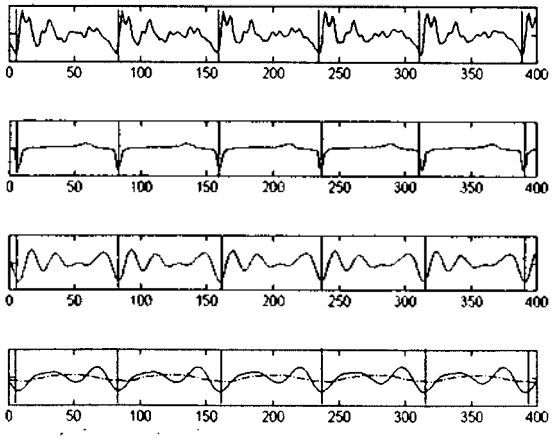


그림 6. 안정된 유성음 구간의 예
(speech/DEGG/scale3/scale4(-),scale5(-.))

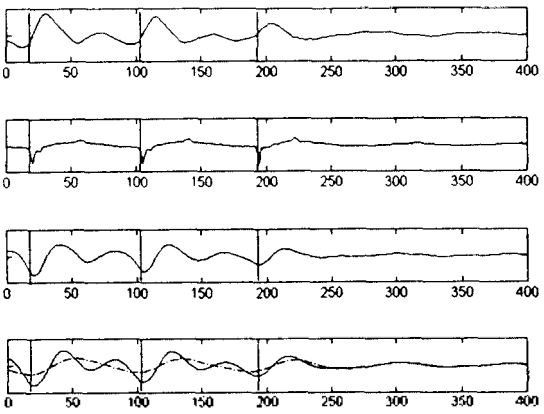


그림 7. voice offset 구간의 예
(speech/DEGG/scale3/scale4(-),scale5(-.))

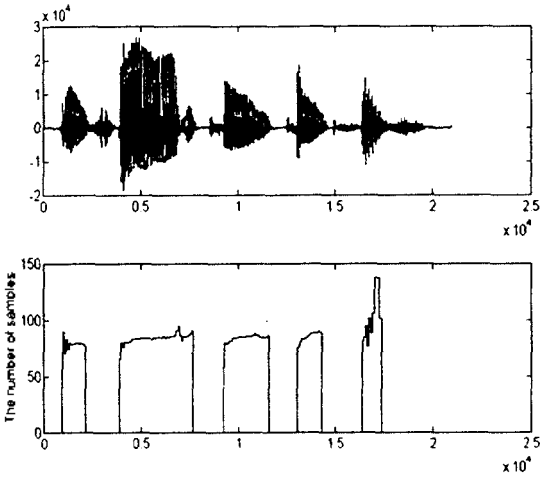


그림 8. 문장 A의 음성신호 및 피치 contour

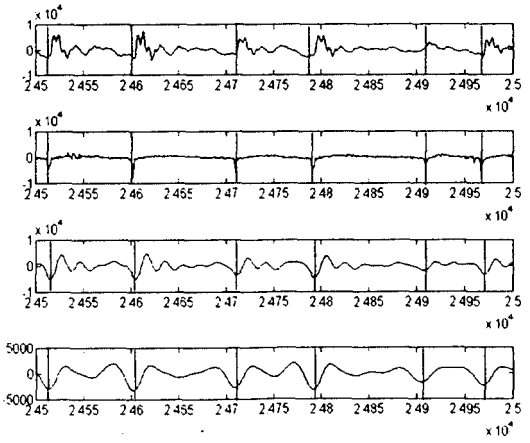


그림 9. 피치주기가 non-stationary한 구간의 예
(speech/DEGG/scale3/scale4)

V. 결 론

본 논문에서는 Mallat의 quadratic spline 웨이브렛 함수를 이용한 음성신호의 웨이브렛 변환으로 epoch을 검출하고 검출된 epoch을 이용하여 피치검출 알고리즘을 제안하였다. 검출된 epoch을 DEGG와 비교하였을 때 대체로 잘 일치함을 보였으나, 화자에 따라서 스케일 3, 4, 5에서의 local maxima 위치값에 많은 차이를 보였다. 이러한 variation에 대한 연구를 위해서 더 많은 음성신호와 DEGG에 대한 분석이 필요하다. 또한, 웨이브렛 변환을 이용한 epoch의 검출에서 threshold값의 결정 문제가 있는데 threshold값의 결정에는 녹음 level이나 화자에 따른 변화 등을 고려해야 함으로 일정한 threshold값을 적용시키기가 어렵다. 따라서 정확한 피치를 검출하기 위해서는 epoch 검출시에 적응적으로 threshold를 정해주는 방법과 epoch 검출후의 error를 보상해주기 위한 후처리방안에 대한 연구가 필요하다.

본 연구는 한국과학재단의 핵심전문연구비 (과제번호: 971-0917-103-2) 지원으로 수행되었으며, 지원에 감사드립니다.

참 고 문 헌

1. 신우용, 김정찬, 배건성 "2-채널(음성 및 EGG) 신호분석에의한 피치검출," *한국음향학회*, vol.15, no.5 1996
2. Shubha Kadambe and G. Faye Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals," *IEEE Trans. Information Theory*, vol. 38, no. 2, Mar.1992.
3. Glenn A. Shelby, "Tone detection using wavelet transforms," University of Alabama in Huntsville, Jul. 1995.
4. DU Limin, and HOU Ziqiang, "Determination of the Instants of Glottal Closure from Speech Wave Using Wavelet Transform," *ICSPAT*,vol.1, Oct.1996.
5. S.Mallat and W. L. Hwang, "Singularity Detection and Processing with Wavelets," *IEEE Trans. Information Theory*, vol.38, no.2, Mar. 1992.
6. D.G. Childers and A. K. Krishnamurthy, " A critical review of electroglottography," *CRC Crit. Rev. Bioeng.*, Vol. 12, no. 2, pp.131-164, 1985.
7. D.G. Childers, A.M. Smith, and G.P. Moore, "Relationships between electroglottography, speech and vocal cord," *Folia Phoniatrica*, Vol. 36, pp. 105-108 , 1989.
8. A.M. Smith and D.G. Childers, "Laryngeal evalution using features from speech and the electrograph," *IEEE Trans. Biomed. Eng.*, Vol. 30, pp. 755-759, Nov. 1983.