1995年度 定期總會 및 秋季學術大會 論文集 1995.11.18

웨이브렛 변환을 이용한 음성 신호의 피치 검출

이민우, 손준일, 최동우, 백승화, 김진수
명지대학교 전기공학과

Pitch Detection of Speech Signals Using Wavelet Transform

Min-Woo Lee, Joon-Il Sohn, Dong-Woo Choi, Seung-Hwa Beack, Jin-Soo Kim
Department of Electrical Engineering MyongJi Univ.

ABSTRACT

In this paper, wavelet transform with multi-resolution property is used to improve the accuracy of pitch estimation of speech signal. Pitch detection of speech signal is based on the local maxima by using wavelet transform. The wavelet transform of a signal is a multiscale decomposition that is well localized in space and frequency. The proposed pitch detection algorithm is suitable for both low-pitched and high-pitched speakers.

1. 서 론

음성신호 처리기를 설계하는데 있어 피치(pitch)는 음성의 분석이나 합성, 인식등에 사용되는 포만트(formant), 강도(intensity)와 더불어 중요한 음성 정보 중 하나이다. 이러한 중요성과 더불어 피치 주기를 추정하는 작업이 어려운 이유는 사람의 성도(vocal tract)가 매우 유동적이고, 그것의 특징이 사람마다 다르기 때문이다. 또한 같은 화자라도 피치 주기가 화자의 감정 상태에 따라 변할 수 있다. 따라서, 음성신호의 피치를 검출할 때에 고려해야 할 사항은 1) 화자의 남녀노소에 영향을 받지 않아야 하고, 2) 화자의 감정이나 개성을 잘 나타내도록 추정범위가 넓어야 한다. 3) 잡음이나 전송 채널의 배경하에서도 실용적인 결과를 얻을 수 있어야 한다. 4) 음성신호의 내용에 무관한 검출 능력을 가져야 한다. 즉, 유성자음으로 시작하는 구간이나 파열음이 연결된 유성음 구간, 비음이 끼어있는 유성음 구간에서도 검출이 잘 되어야 한다.

피치 검출에 대한 연구는 시간 영역, 주파수 영역, 시간-주파수 영역으로 분리하여 처리하고 있는데 기존의 피치 검출 알고리즘들은 길이가 고정된 윈도우의 사용에 의해 얻어지는 음성신호의 각각의 세그먼트 안에서 안정적(stationary)이고, 각 세그먼트는 적어도 두개의 완전한 피치 주기를 포함한다는 가정하에 평균 피치 주기를 계산한다. 그러므로 이러한 피치 검출기의 단점은 세그먼트 길이 전체에 대한 피치 주기의 비정상 신호 변화에 대해 민감하지 않고 낮은 피치와 높은 피치를 가진 화자들에게 부적당하다는 것이다.

본 연구에서는 음성신호의 피치 간격을 추정하기 위해 웨이브렛 변환에 기반한 알고리즘을 제안하였다. 웨이브렛 변환은 시간-주파수 분석에서 매우 뛰어난 기능을 보여준다. 신호들은 시간과 주파수 영역에서 기본 블록들로 적당하게 나누어짐으로써, 신호의 지역성(localization)을 잘 표현할 수 있다. 이러한 특징은 음성신호를 다른 잡음들과 구별하는데 사용된다. 본 논문에서는 음성 신호의 피치 특성을 추출하는데 dyadic 웨이브렛 변환을 사용하였다. 서로 다른 스케일들에서 웨이브렛 크기가 갖는 구역 최대값은 음성신호가 급격하게 변화하는 위치를 찾는데 사용된다. 본 연구에서는 먼저 /아/, /에/, /이/, /오/, /우/, /애/, /어/, /으/ 8개의 단모음의 피치를 찾고, /가/, /나/, /다/, /라/ 4개의 자모 합성음의 특징을 찾는데 주안점을 두었다.

2. 웨이브렛 변환 (Wavelet Transform)

신호변환의 목적은 신호로부터 특징 정보를 추출하여 신호 분석을 쉽게 하고자 함에 있다. 임의의 신호의 주파수 특성을 쉽게 알 수 있는 변환의 방법으로는 푸리에 변환이 널리 사용되고 있다. 그러나 푸리에 변환의 전제조건은 입력 신호의 특성이 시간상으로 변하지 않는 정상신호라는 가정에서부터 시작되기 때문에 비정상 신호의 분석에는 적합하지 못하게 된다. 이에 비정상 신호의 분석에 적합한 변환방법으로 웨이브렛 변환을 사용할 수 있다. 웨이브렛 변환이란 웨이브렛(wavelets)이라고 불리는 작은 파형들을 이용하여 신호나 시스템, 프로세스 등을 모델화하는데 사용하는 수학적인 방식이다. 작은

파형, 웨이브렛(wavelet)이 될 수 있는 조건은 진동하여야 하고, 시간축의 양 방향에 대해 급격히 0으로 줄어들어야 한다. 신호 $x(t)$의 dyadic 웨이브렛 변환($DyWT$)은 다음과 같다.

$$D_yWT_x(b,2^i) = \frac{1}{2^i}\int_{-\infty}^{\infty} x(t)g(\frac{t-b}{2^i})dt \quad \cdots\cdots (1)$$
$$= x(t) \otimes g_{2^i}(t)$$

스케일 파라미터, $a=2^i$를 사용하여 웨이브렛 변환을 계산하였다. 여기서, $g(t)$는 다음과 같은 조건을 만족하는 웨이브렛 함수 $g(t)$의 공액 복소수이다.

$g_{2^i}(t)=\frac{1}{2^i}g(\frac{t}{2^i})$, (⊗는 컨벌루션 연산자를 나타낸다)
신호처리의 관점에서 $DyWT$는 Constant-Q, 옥타브 대역, 또는 임펄스 응답이 $\frac{1}{2^i}g(\frac{1}{2^i})$인 대역통과 필터의 출력으로 생각할 수 있다. 그러한 필터들 각각의 대역폭과 중심 주파수는 $\frac{1}{2^i}$에 비례한다.

웨이브렛을 사용하여 이산화된 음성 신호를 미분하게 되면, 잡음 성분이 증폭되는 문제가 발생한다. 이러한 문제점을 해결하기 위하여 미분을 하기 전에 잡음 성분을 걸러주는 여과 과정이 필요한데 이러한 과정을 smoothing이라 한다. 이러한 과정은 웨이브렛 변환에 의해 수행할 수 있다. 먼저, 음성 신호의 잡음 성분을 제거하기 위해 여과를 할 수 있는 웨이브렛(wavelet)을 선택한다. 다음으로 각 스케일에서 음성신호의 웨이브렛 변환은 smoothing 정도를 스케일 변수화하여 음성신호의 고주파 성분을 제거한다.

Smoothing 함수 $\theta(t)$가 일차 미분인 웨이브렛함수 $g(t)$를 선택한다면, $DyWT$의 구역 최대값은 느린 변화(slow variance)를 보인다. 여기서, smoothing 함수란 그것의 푸리에 변환이 저주파수 영역에서 에너지 집중을 갖는 함수이다. 그러므로 smoothing 함수의 일차 미분인 웨이브렛을 사용하는 $DyWT$의 구역 최대값은 급격한 변화나 성문 폐쇄에 의해 발생하는 음성 신호에서의 과도 성분을 검출하는데 유용하다. Mallat은 몇개의 연속적인 dyadic 스케일을 통해 시간 $t=t_0$에서 $DyWT$의 구역 최대값임을 증명하였다. 그는 세개의 연속적인 dyadic 스케일들에 대해 두개를 비교하여 $DyWT$ 구역 최대값의 상관성을 이용한 효율적인 영상 코딩 알고리즘을 개발하였다. 본 연구에서는 두개의 연속적인 스케일들을 통해 $DyWT$ 구역 최대값의 상관성에 대해 조사하였다.

3. 웨이브렛 변환과 피치 검출

웨이브렛 변환은 주파수와 시간 영역에 대해 모두 좋은 지역성(위치와 정보를 포함함)을 갖는다. 주파수 영역에서의 지역성이란 웨이브렛 변환의 스케일과 주파수 대역간의 일치함을 의미한다. 그러므로 웨이브렛 변환을 다중주파수 분해라고도 한다. 시간 영역에서의 지역성은 웨이브렛 변환이 푸리에 변환보다 더 뛰어난 주요 장점이다. 푸리에 변환은 전체 신호를 균등하게 볼 수 있도록 한 반면, 웨이브렛 변환은 각 위치에서 신호의 구역 균등성을 볼 수 있도록 한다. 음성신호의 크기(magnitude)는 변화를 측정할 수 있기 때문에, 이러한 크기의 구간 최대값은 급격한 변화와 일치한다. 크기의 구간 최소값은 급격한 변화가 아닌 느린 변화에 대해 일치한다. 각구간의 최대값에 대해, 그것의 위치와 값이 기록된다. 구간 최대값의 표현은 특히 음성신호의 피치 주기를 검출하는데 있어 뛰어난 정보를 포함한다. 만약 신호 x(t)나 그것의 미분들이 불연속들을 갖는다면, x(t)의 $DyWT$의 modulus, $|DyWT_x(b,2^i)|$는 불연속점들 주위에 구역 최대값을 나타낸다. 이러한 성질은 성문이 폐쇄될때 공기흐름을 급격하게 변화시키기 때문에 피치를 검출하는데 큰 역할을 한다.

음성 신호의 피치에 대해 구역 최대값이 갖는 의미는 다음과 같다.

1. 구역 최대값(local maxima)의 위치는 피치의 위치이다.
2. 구역 최대값의 크기는 피치의 세기를 나타낸다.
3. 구역 최대값의 부호는 피치점에서 신호값들이 올라가는지 또는 내려가는지를 가리킨다.
4. 스케일들간의 구간 최대값을 늘어 놓는 것은 피치의 형태를 특징지어준다.

음성 신호에서 한 세그먼트의 dyadic 웨이브렛 변환은 스케일 $a=2^i$, $i=i_s, i_{s+1}, \cdots\cdots, i_f$에 의해 계산된다. 여기서 i_s은 시작 스케일을 i_f은 마지막 스케일을 나타낸다. 각각의 스케일 2^i에서 $DyWT(b,2^i)$의 쉬프트 인자 b로 표현되는 구역 최대값의 위치를 찾아낸다. 각 스케일별로 나타난 신호들간에 구역 최대값(local maxima) 위치를 비교하였을 때 서로 일치한다면, 이러한 최대값의 위치들은 성문 폐쇄에 의해 발생하는 과도 현상의 시간과 같다고 볼 수 있기 때문에, 두개의 구역 최대값들 사이의 시간격을 측정하는 것에 의해 피치 주기를 계산한다.

본 연구에서는 스케일 파라미터의 개수를 3개로 제한하여 계산상 시간 절감을 할 수 있었는데, 이는 음성이 보통 10 옥타브에 걸쳐 있고, 유성음 신호의 피치나 기본 주파수는 저주파수(30-500Hz)현상인데 반하여, 무성음들은 랜덤한 고주파수 현상이라는 사실에 의해 단지 3

개의 dyadic 스케일들에서 $DyWT$을 계산하는 것만으로도 피치 주기를 추정하는데 충분하다는 것을 알 수 있다. 세개의 dyadic 스케일들은 다음과 같이 선택되었다. 입력 중심주파수 f_{c_i}와 입력 대역폭 Δf_i를 갖는 웨이브렛이 주어진다면, 방정식 $a=\frac{f_{c_i}}{f_{c_o}}$ ··· (3) 을 사용하여 필요한 출력 중심주파수 f_{c_o}와 일치하는 스케일 파라미터 'a'를 선택할 수 있다. 본 연구에서는 웨이브렛의 입력 대역폭 $\Delta f_i=2\times f_{c_i}$와 출력 대역폭 $\Delta f_o=2\times f_{c_o}$를 선택하였다. 만일, 방정식 (3)에서 $\frac{f_{c_i}}{f_{c_o}}$가 2의 임의의 급수와 같지 않다면 가장 가까운 급수로 반올림을 한다.

여기서는 입력 중심주파수 $f_{c_i}=8000\ Hz$이고, 입력대역폭 $\Delta f_i=16000\ Hz$인 cubic spline 웨이브렛을 발생시켰다. 먼저, 하위 대역으로 $a=2^{i_0}$을 필요한 출력 주파수 $f_{c_o}=1000\ Hz$, 대역폭 $\Delta f_{c_o}=2000\ Hz$으로 선택하는 것에 의해 스케일 파라미터를 구한다. 다음으로 상위 대역 $a=2^{i_0}$를 기본 주파수의 영역과 주파수들이 일치하도록 하기 위해 필요한 출력 중심주파수 $f_{c_o}=250\ Hz$이고, 대역폭 $\Delta f_{c_o}=500\ Hz$인 것을 선택하는 것에 의해 스케일 파라미터를 구한다.(30~50 Hz는 결코 걸러지지 않는다.) 다음으로 가장 낮은 스케일에서 시작하는 $DyWT$를 계산하고, 가장 높은 스케일 파라미터에 도달할 때까지 스케일 파라미터를 계속 배로 증가시킨다. 본 연구에서는 스케일 파라미터 상의 가장 낮은 대역와 가장 높은 대역으로 각각 $a=2^3$과 $a=2^5$로 놓았다. 단지 세개의 스케일들에서 $DyWT$를 계산하였다. 이것은 $DyWT$의 계산상의 복잡성을 확실하게 감소시킨다. 제안한 알고리즘에서 무성음의 주파수는 기본 주파수 영역보다 매우 높기 때문에 더 높은 스케일들에서는 무성음이 걸러진다. 그런 까닭에, 두 스케일들을 비교하여 $DyWT$의 구역 최대값이 상관성이 있는지 검사할 뿐 아니라 음성신호의 주어진 세그먼트에서 임의의 문턱값를 가지고서 $DyWT$의 최대 진폭을 비교하는 것에 의해 유성음과 무성음으로 분류할 수 있다.

4. 실험 및 결과

8비트 8KHz로 샘플링된 /아/, /에/, /이/, /오/, /우/, /어/, /애/, /으/등 8개 단모음에서 512개의 샘플들을 취해 dyadic 웨이브렛 변환을 사용하여 스케일별로 분석하였다. 그리고 자모 합성음 /가/, /나/, /다/, /라/에 대한 웨이브렛의 extrema를 구하였다. 본 논문에서는 단모음 /오/음과 자모 합성음 /가/음에 대해서 실험하였다.

· 단모음 /오/의 피치 검출

먼저, 8개 단모음들 중 /오/음의 피치 주기를 추정하기 위해 D_yWT 방법과 기존의 피치 추출 방법인 자기상관계수법을 사용하였다. 그림 1 은 단모음 /오/음의 원파형이다. 샘플들 단위는 ms로서 8000Hz로 1초간 샘플링되었다. 그림 2는 자기상관계수법을 적용한 것이고, 그림 3은 스케일 $a=2^3,2^4,2^5$에서 각각 계산되어진 신호 /오/의 D_yWT을 나타낸 것이다.

신호 $x(t)$의 자기상관계수법 $R_x(\tau)$는,

$$R_x(\tau) = \int_{-\infty}^{\infty} x^*(t)x(t+\tau)dt \text{ 로서 정의된다.}$$

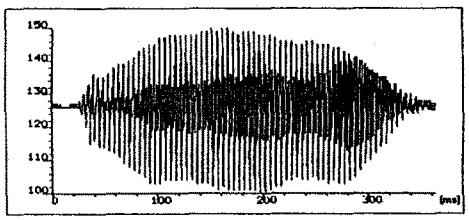

그림 1. 단모음 /오/음의 원파형

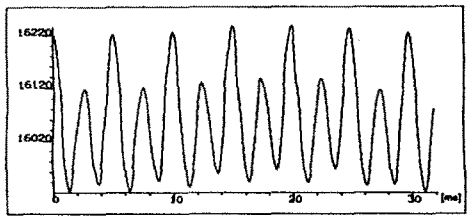

그림 2. 단모음 /오/음의 자기상관계수법을 실행한 결과

주기 신호의 자기상관함수는 신호의 주기성을 보인다. 많은 음성 신호처리 분야에서 음성 신호의 피치 주기를 계산하는데 자기상관함수의 성질을 이용한다.

그림 3은 비록 3개의 스케일 영역에 대해서 나타냈지만 스케일 파라미터가 증가함에 따라 고주파수 정보가 걸러짐을 볼 수 있다. 이에 따라, 피치 주기를 정밀하게 계산하기 위해 2^4 또는 2^5의 스케일에서 계산되는 D_yWT를 선택할 필요가 있다.

그림 4는 local extrema를 나타내었다. 각 스케일들을 비교하여보면 최대값의 위치가 일치하는 것을 쉽게 볼 수 있다. /오/음의 피치 길이는 5 ms이다.

알고리즘은 스케일 $a=2^4$에서 계산된 D_yWT을 선택하였다. 왜냐하면 스케일 $a=2^4$와 $a=2^5$에서 계산된 D_yWT의 구역 최대값이 문턱값(threshold)의 위치와 일치하기 때문이다. 피치 주기 계산의 정확성은 웨이브렛 함수로서 어떠한 것을 선택했는가에 따라 다르다.

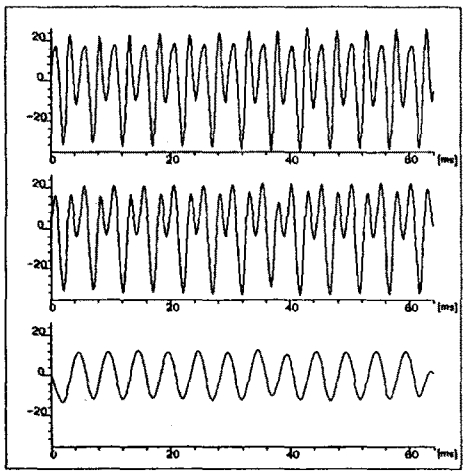

그림 3. 스케일 $a=2^3$, $a=2^4$, $a=2^5$에서 각각 계산된 /오/음의 D_2WT

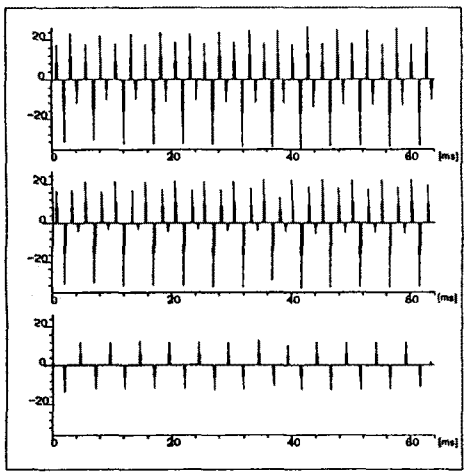

그림 4. 스케일 $a=2^3$, $a=2^4$, $a=2^5$(위로부터)에서 각각 계산된 local extrema

자모 합성어 /가/음의 분석

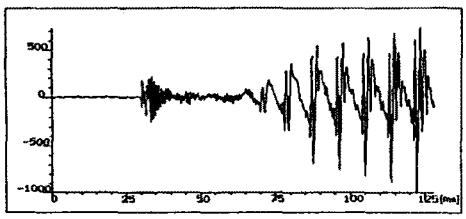

그림 5. 자모 합성어 /가/음의 원신호

모음과는 달리 파열음('ㄱ', 'ㄷ', 'ㅂ')은 스펙트럼의 시간적인 변화가 매우 심하고, 마찰음('ㅅ', 'ㅍ'등)은 특별한 포먼트 주파수를 찾아 볼 수 없는 백색 잡음의 형태를 가지고 있다. 그 외에도 대부분의 자음은 모음에 비해 특징이 분명하지 않다. 따라서 여기서는 자음과 모음이 결합된 형태의 음절들 중 /가/음에 대해서 나타내어 보았다. 웨이브렛 변환을 하여 각 스케일별로 extrema를 나타내 보였지만 모음부에 대해서만 확실한 구역 최대값을 표현하고, 자음부에 대해서는 어떠한 특징점을 찾기가 어렵다.

자기상관계수법에 있어서 세그먼트 길이의 선택은 매우 중요하다. 이 방법은 길이 L인 신호 세그먼트의 평균 피치 주기를 추정하고, 이러한 것은 적어도 두개의 피치 주기가 선택된 세그먼트 안에 필요하다. 만약 음성 세그먼트가 너무 짧다면, 알고리즘은 피치주기를 정확하게 추정할 수 없다. 반대로 세그먼트가 너무 길다면 이러한 알고리즘들은 주기에서 주기까지의 피치 주기 길이에서 나타나는 비정상 변화들을 감지 할 수 없게 된다. 그러나, D_2WT의 경우에 있어서는 성문 폐쇄시의 시간 간격에 의해 피치 주기를 검출하기 때문에 세그먼트 길이에는 별 영향을 받지 않는다.

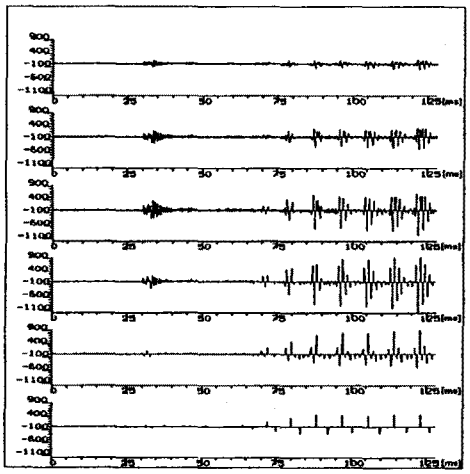

그림 6. 자모 합성어 /가/음을 스케일 6으로 D_2WT한 후, 각 스케일별로 local extrema를 나타낸것

5. 결론

본 연구에서는 웨이브렛을 이용하여 보다 정확한 음성 신호의 피치를 검출하고자 하였다. Dyadic 웨이브렛 변환을 이용한 피치 검출 알고리즘의 장점은 분석하고자 하는 윈도우 안의 신호 상태에 영향받지 않고, 일정 크기의 잡음하의 신호에 대해서도 피치 주기를 매우 정확하게 추정한다. 넓은 영역의 피치 주기들에 대해 적당하다. 또한 피치 주기의 시작과 음성 신호의 주어진 세그먼트에 존재하는 피치 주기의 개수를 감지할 수 있다. 전체 스케일 계산에 대해 단지 두개나 세개의 스케일들에 대해서만 D_2WT을 계산함에 따라 간단하게 처리할 수 있다.

참 고 문 헌

[1] Olivier Rioul and Martin Vetterli, "Wavelets and Signal Processing," IEEE SP Magazine, pp.14-38, October 1991.

[2] Silvio Montrésor, Marc Baudry, "Pitch estimation of speech signal with the wavelet transform," pp.2017-2020.

[3] Shubha Kadambe and G. Faye Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals," IEEE Transactions on Information Theory, Vol.38, No.2, pp.917-924, March 1992.

[4] E. Ambikairajah, M. Keane, L. Kilmartin and G. Tattersall, "The Application of the Wavelet Transform for Speech Processing,"

[5] 박 경범, "선형예측분석법에 의한 음성의 압축과 재생," 하늘소 출판사, pp.31-132.

[6] Cuiwei Li, Chongxun Zheng, and Changfeng Tai, "Detection of ECG Characteristic Points Using Wavelet Transforms", IEEE Transactions on Biomrdical Engineering, Vol.42, No.1, pp.21-28, January 1995.

[7] Olivier Bertrand, Jorge Bohorquez, and Jacques Pernier, "Time-Frequency Digital Filtering Based on an Invertible Wavelet Transform: An Application to Evoked Potentials", IEEE Transactions on Biomrdical Engineering, Vol.41, No.1, pp.77-88, January 1994.

[8] Sifen Zhong, "Edge Representation from Wavelet Transform Maxima", tech. rep., Department of Computer Science, New York University, September