시간-주파수 혼성 피치검출기의 성능개선에 관한 연구

조왕래, 김종국, 배명진
숭실대학교 정보통신공학과

A Study On a Pitch Detection in Time-Frequency Hybrid Domain

Wang-rae Jo, Jong-kuk Kim, Myung-jin Bae
Information and Telecommunication Engr., Soongsil University
wrjo@unitel.co.kr, mjbae@ssu.ac.kr

요 약

본 논문에서는 시간-주파수 혼성 영역 피치 검출법을 제안하였다. 음성신호를 주파수 영역으로 변환하고 주파수 영역에서 위상 성분을 조절하여 시간영역으로 역변환 함으로써 피치 피크가 최대가 되도록 하여 용이한 피치검출이 가능하였다. 또한 처리시간을 단축하기 위하여 FFT와 IFFT의 비트 재정렬을 생략하여 처리할 수 있는 방법을 제안하였다. 성능 측정 결과 기존의 캡스트럼 검출법에 비하여 검출성능은 우수하면서도 처리시간은 84.8%로 단축됨을 알 수 있었다.

1. 서 론

음성 신호 처리 분야에 있어 기본주파수 검출은 매우 중요하다. 음성신호의 기본주파수를 정확하게 검출할 수 있다면 화자의 영향이 배제된 음성인식이 가능해져 인식의 정확도를 높일 수 있으며, 피치 변경도 용이해져 합성음의 개성을 쉽게 변경할 수 있게 된다.

이러한 중요성 때문에 피치검출에 대한 다양한 방법들이 제안되었으며, 이들은 처리영역에 따라 시간 영역법, 주파수 영역법, 시간-주파수 혼성 영역법으로 나눌 수 있다. 시간 영역법은 ACM법, AMDF법, 병렬처리법 등이 있으며 처리법이 매우 간단하지만 천이구간에서의 피치검출이 어렵다는 단점이 있다[1][2]. 주파수 영역법은 고조파 분석법, Lifter법, Comb-filtering법 등이 있다. 주파수 영역법은 음소의 천이나 변동에 영향을 적게 받는 장점이 있지만 주파수 정확도를 높이기 위해서는 프레임 사이즈가 커지므로 처리시간이 길어지고 변화특성에 둔감해지는 단점이 있다. 시간-주파수 혼성 영역법은 시간영역법과 주파수영역법의 장점을 취한 방법이지만 영역변환에 필요한 계산과정이 복잡하다는 단점이 있다[3].

본 논문에서는 시간-주파수 혼성 영역 피치 검출법을 제안하였다. 음성신호를 주파수 영역으로 변환하고 주파수 영역에서 위상 성분을 조절하여 시간영역으로 역변환 함으로써 피치 피크가 최대가 되도록 하였다. 따라서 피치 검출이 정확하게 이루어 질 수 있다. 또한 혼성영역법의 단점인 영역변환 처리 시간을 단축하기 위하여 영역변환을 위한 FFT와 IFFT연산에서 비트-재정렬 과정을 생략하는 방법을 제안하였다.

2. 호모몰픽 디컨벌루션

음성신호 분석의 기본적인 가정중의 하나는 음성신호는 시간에 따라 느리게 변화하는 선형 시변 시스템의 출력으로 표현할 수 있다는 것이다. 이것은 음성신호의 짧은 구간만을 고려할 때 각 세그먼트는 준주기적인 임펄스나 불규칙 잡음에 의해 여기된 선형 시불변 시스템의 임펄스 응답으로 모델링된다는 것이다. 음성신호 분석은 컨벌루션된 여기성분과 성도성분을 분리하여 파라미터화하는 것을 말한다. 호모몰픽 디컨벌루션은 음성의 이러한 특성을 이용하여 여기성분과 성도성분을 분리하는 기법으로 호모몰픽 필터링이라고도 한다[2].

호모몰픽 필터는 시스템을 통과하는 동안 원하지 않는 성분은 제거하는 반면 원하는 성분에는 영향을 미치지 않는다. 컨벌루션된 신호를 분리하고 복원하기 위한 일반적인 호모몰픽 시스템은 그림 2-1에 나타낸 바와 같이 세 개의 호모몰픽 시스템의 직렬접속으로 표현할 수 있다.

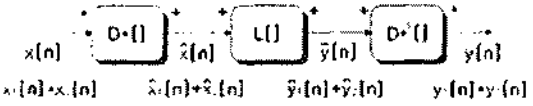

그림 2-1. 호모몰픽 디컨벌루션 시스템

그림 1에서 첫 번째 시스템은 특성 시스템이라 하며, 식 (1)과 같이 컨벌루션 입력을 취하여 각 입력에 대응하는 출력의 합으로 출력한다[3].

$$\begin{aligned} D_*[x(n)] &= D_*[x_1(n) * x_2(n)] \\ &= D_*[x_1(n)] + D_*[x_2(n)] \\ &= \hat{x}_1(n) + \hat{x}_2(n) = \hat{x}(n) \end{aligned} \tag{1}$$

이러한 특성 시스템은 컨벌루션을 곱의 형태로 변환하는 Z변환과 곱을 합의 형태로 변환하는 로그연산의 특성을 이용하여 그림 2-2와 같이 구현할 수 있다.

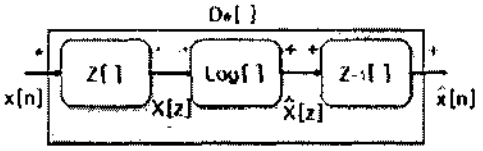

그림 2-2. 호모몰픽 디컨벌루션의 특성 시스템

특성 시스템의 입력이 여기신호 $s(n)$ 과 성도성분 $h(n)$ 의 컨벌루션이라 한다면 입력신호 $x(n)$ 은 식 (2)와 같이 표시할 수 있고 Z변환은 식 (3)과 같이 표현할 수 있다.

$$x(n) = s(n) * h(n) \tag{2}$$

$$X(z) = S(z) \cdot H(z) \tag{3}$$

이것은 로그연산에 의해 합의 형태로 변환되고 역 Z변환에 의해 시간영역으로 변환된다.

$$\begin{aligned} \hat{X}(z) &= \log[X(z)] \\ &= \log[S(z) \cdot H(z)] \\ &= \log[S(z)] + \log[H(z)] \\ &= \hat{S}(z) + \hat{H}(z) \end{aligned} \tag{4}$$

$$\hat{x}(n) = \hat{s}(n) + \hat{h}(n) \tag{5}$$

이러한 특성 시스템의 출력을 복소 켑스트럼이라 하고 여기성분과 성도성분의 합으로 표현된다.

3. 하이브리드 피치검출법의 성능개선

3.1 위상조절에 의한 검출성능 개선

어떤 신호가 서로 다른 주파수를 갖는 세 개의 신호로 구성되어 있다고 하더라도 그림 3-1(a)와 같이 그 위상이 서로 다르다면 합성파형은 그림 3-1(b)와 같이 한 주기 파형의 변화 양상이 복잡한 형태로 나타나게 된다. 이렇게 복잡한 변화양상은 켑스트럼 영역에서의 피크크기를 감소시켜 정확한 피치 검출을 매우 어렵게 만든다. 그러나 그림 3-1(c)와 같이 세

신호의 위상을 조절하여 동일위상의 코사인 신호로 합성한다면 그림 3-1(d)와 같이 세 신호의 주기가 일치하는 지점에서 특히 강조된 피크가 나타나게 되어 합성신호의 기본 주기를 검출하기가 매우 용이하게 된다.

음성신호의 경우도 주파수와 위상이 다른 여러 신호의 조합으로써 서로 다른 위상으로 인해 복잡한 변화양상을 가지며 이로 인해 피치검출이 어려워진다. 그러나 앞의 예와 같이 주파수 영역으로 변환하고 위상을 동위상으로 변환하여 역변환하면 여러 신호의 주기가 일치하는 기본 주파수 지점에서 강조된 피크가 나타나게 된다. 음성신호의 경우 일반적인 피치범위가 2.5ms에서 25ms 사이이므로 11kHz로 샘플링 된 신호의 경우 27~270표본 사이에서 최대값을 검출하면 간단히 피치를 검출할 수 있다. 이를 블록도로 표현하면 그림 3-2와 같다.

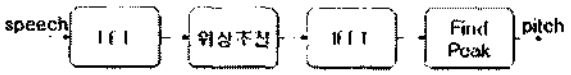

그림 3-2. 혼성영역 피치검출과정

3.2 비트재정렬 생략에 의한 검출속도 개선

앞에서 제안한 피치 검출법은 음성신호를 주파수 영역으로 변환하고 위상을 조절한 후 다시 시간영역으로 변환하여야 하며 이를 위하여 FFT와 IFFT 연산이 사용된다.

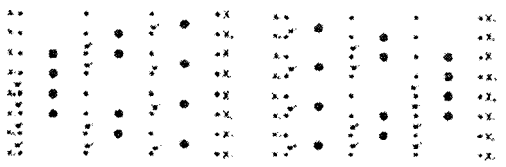

그림 3-3. 8-point FFT의 흐름도

(a) 정상순서 입력의 DIT 흐름도

(b) 비트-재정렬 입력의 DIT 흐름도

이러한 FFT, IFFT 연산은 연산의 특성상 계산하고자 하는 데이터 샘플수가 $N = 2^{\nu}$ (ν 는 정수)이 되어야 한다는 것과 그림 3-3에 나타낸 바와 같이 입력배열과 출력배열의 순서가 서로 일치하지 않으므로 FFT 수행 전이나 수행 후에 배열의 순서를 재정렬해 주는 비트-재정렬(bit-reversing)이 필요하며 FFT 계산량에 큰 오버헤드로 작용하게 된다. 이러한 오버헤드는 켑스트럼 분석과 같이 시간-주파수 영역 변환이 잦은 연산의 처리 속도에 큰 영향을 미치게 된다[7].

본 논문에서는 음성신호의 혼성 영역에서의 피치를 검출하고자 할 때 FFT와 IFFT의 비트-재정렬 과정을 생략함으로써 처리시간을 단축하는 방법을 사용하였다. 기존의 혼성 영역 처리에서는 FFT와 IFFT에 동일한 알고리즘을 사용함으로써 필연적으로 비트-재정렬을 수행하여야 하였으며, 이러한 오버헤드는 처리시간에 큰 영향을 주었다. 그러나 FFT와 IFFT에 다른 방법을 사용한다면 비트-재정렬 과정을 생략할 수 있게 된다. 즉, FFT에는 정상순서 입력의 DIT 방법을 사용하고 그 결과를 비트-재정렬된 입력의 DIT 방법을 사용하여 IFFT하면 정상순서의 출력을 얻을 수 있게 된다. 기존의 처리과정과 제안한 처리과정을 그림 3-4에 비교하여 나타내었다.

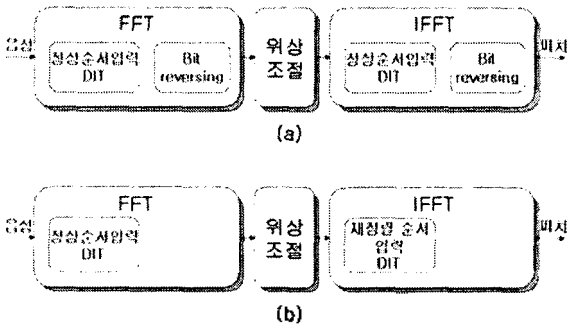

그림 3-4. 혼성 영역 처리시간 단축법

(a) 기존의 방법

(b) 제안한 방법

4. 실험 및 결과

개선된 혼성 영역 피치 검출법의 성능을 평가하기 위하여 기존의 켑스트럼 피치 검출법과 제안한 피치

검출법을 IBM-PC/Pentium IV에서 C++로 구현하였다.

피치 검출의 정확성을 비교하기 위하여 기존의 켑스트럼 피치검출법과 제안한 방법으로 발성문장 "안수네 꼬마는 천재소년을 좋아한다."의 피치 변화도를 측정하였다. 그림 4.1(b)에는 유성음 구간에서 기존의 켑스트럼 검출법을 이용하여 구한 피치변화도를 나타내었으며 4.1(c)에는 제안한 방법으로 구한 피치 변화도를 나타내었다. 그림 4.2에는 발성문장 중의 한 프레임의 파형과 제안한 방법에 의해 위상조절된 음성 프레임을 나타내었다. 기존의 방법에 비하여 제안한 방법의 피크 피치가 크게 두드러져 피치를 검출하기가 용이함을 알 수 있다.

기존의 방법과 제안한 방법의 처리 시간을 비교하기 위하여 프레임의 크기를 128, 256, 512 샘플로 변화시켜 가면서 전체 문장의 처리 시간을 측정하였다. 표 4.1에 나타낸 바와 같이 512 샘플에 대한 기존의 켑스트럼 연산 시간보다 제안한 방법이 3,550usec 소요되어 기존 방법 4,188usec의 84.8%로 단축됨을 알 수 있었다.

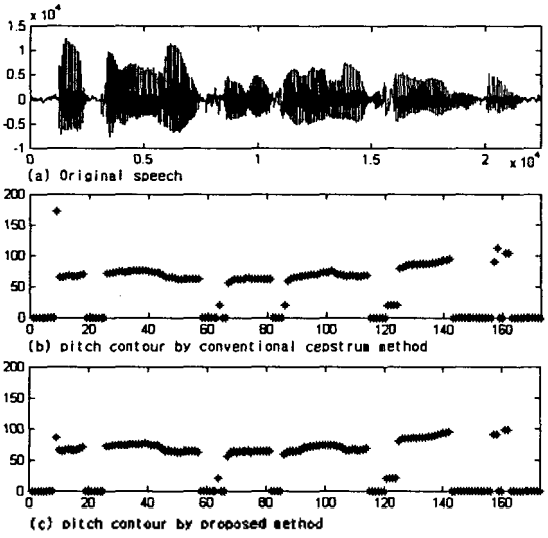

그림 4-1. 발성문장의 피치 변화도

(a) 발성문장의 파형

(b) 기존의 방법에 의한 피치변화도

(c) 제안한 방법에 의한 피치변화도

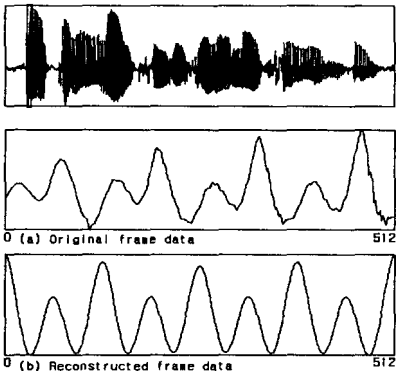

그림 4-2. 위상 조절된 프레임 데이터

표 4.2 처리 시간의 비교

	Processing Time [us]		Rate (B/A)
	Conventional method (A)	Proposed method (B)	
128 points	835	712	85.3%
256 points	1,863	1,638	87.9%
512 points	4,188	3,550	84.8%

5. 결 론

본 논문에서는 시간-주파수 혼성 영역 피치 검출법을 제안하였다. 음성신호를 FFT하여 주파수 영역으로 변환하고 주파수 영역에서 위상 성분을 조절하여 시간영역으로 IFFT함으로써 피치 피크가 최대가 되도록 하였다. 이렇게 함으로써 용이한 피치검출이 가능해 진다. 또한 처리시간을 단축하기 위하여 정상순서 입력 DIT FFT, 비트-재정렬 입력 IFFT를 수행함으로써 기존 방법의 84.4%로 연산시간을 단축할 수 있었다.

참고문헌

1. L.R.Rabiner, R.W.Schafer, Digital Processing of Speech Signals, Prentice-Hall, 1978.
2. P.E.Paparnichalis, Practical Speech Processing, Prentice-Hall, 1987.
3. S.Seneff, "Real Time Harmonic Pitch Detection", IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-26, pp.358-365, Aug. 1988.