청각 모델을 이용한 Cochannel 음성에서의 피치 추출에 관한 연구

速報論文
49D-6-6

A Study on Pitch Detection using Cochlear Model on Cochannel Speech

愼 大 揆* · 愼 重 寅** · 李 宰 赫*** · 韓 科 辰§ · 朴 相 曦§§
(Dae-Kyu Shin · Joong-In Shin · Jae-Hyuk Lee · Doo-Jin Han · Sang-Hui Park)

Abstract - In this paper, a new pitch estimation method is proposed using the Robinson cochlear model. This estimation method is useful in noisy environments and especially very efficient under cochannel in which two speaker voices exist at the same time.

For the one speaker speech, the pitch can be extracted from just the neurogram of the Robinson cochlear model. In this case, as the estimation is performed in time domain, the exact pitch period can be detected though the pitch period is various. But in noisy and cochannel cases, the neurogram has many spurious peaks, so we use the autocorrelators in the neurogram to manifest the period. If the autocorrelators are used for the all delays, the large amount of calculations is necessary. Due to this defect, we propose that the autocorrelators are used for the part of the delays on which energy is concentrated. First of all, the proposed algorithm is applied to the one speaker speech, and later to the cochannel speech. And then the results are compared with the autocorrelation pitch detection method.

Key Words : Robinson cochlear model, Cochannel speech, Pitch estimation

1. 서 론

유성음은 성대가 진동하면서 발생하고, 이렇게 발생한 진동의 기본 주파수나 이로 인해 나타나는 음성신호의 주기성을 피치라고 한다. 이러한 피치는 발음된 음성의 억양(intonation)과 강세(stress)에 대한 주된 음향적 기준이 되고 음소 판단(identification)에 중요한 파라미터가 된다. 대부분의 저-비트율 음성부호화기(vocoder)는 재구성한 음성이 좋은 음질을 갖도록 하기 위해 정확히 피치를 추정할 필요성을 갖는다. 또한 피치 패턴은 화자인식이나 음성합성에서도 유용하게 사용되는 파라미터이다.

이렇게 피치는 음성신호의 분석 및 응용에 있어서 가장 기본적으로 요구되는 요소이므로 오래 전부터 그 추출 방법이 연구되어왔다. 가장 대표적인 것이 정상적(stationary)인 신호의 주기성 검출에 효과가 큰 자기상관법(Autocorrelation method), AMDF(Average Magnitude Differential Function) 법 등과 같은 non-event method가 있으며[1], 최근에 와서는 그 발화 시점(event)의 검출을 통한 피치 추정법으로 웨이브렛(wavelet)과 같은 방법이 연구되고 있다[2]. 자기상관법과 AMDF 법은 간단한 연산을 통하여 주기성을 추출할 수 있으나 고정 윈도우를 사용한 음성 세그먼트에 대하여 평균적인 피치 값을 구하기 때문에 비정상(non-stationary)적인 경우나 피치가 변화하는 경우에는 효과가 떨어진다. 웨이브렛을 이용하여 피치를 추정하는 경우, 성대가 열리는 순간의 시점에서 음성 신호가 크게 변화한다는 사실을 이용하여 이러한 부분을 추정하는 방식으로 비정상적이고 SNR(Signal to Noise Ratio)이 낮은 경우에도 좋은 성능을 나타낸다[2]. 그러나 이와 같은 피치 검출 방법들은 한 화자의 음성에 다른 화자의 음성이 섞인 cochannel의 상황에서는 SIR(Signal to Interference Ratio)에 따라 두 화자중 한 사람의 피치도 검출이 거의 불가능한 상황이 발생한다.

본 연구에서는 청각 모델을 이용하여 음성신호의 피치를 추정하는 방법을 제안하고자 한다. 인간의 귀는 심한 잡음이 섞여 있거나 혹은 피치가 비정상적으로 변화하는 경우에도 정확히 음성을 인식한다. 심지어 여러 사람이 동시에 말하는 상황에서도 인식이 가능하다. 따라서, 청각모델을 사용한다면 이러한 장점을 살리면서 음성 분석이 가능할 것이다. 인간의 청각이 이렇게 여러 피치를 동시에 인지할 수는 있지만, 본 연구에서는 우선 동시에 발생한 두 사람의 음성신호에서 가장 가능성이 높은 한 사람의 음성신호에 대한 피치를 검출하는데 목적을 둔다.

2. 청각모델의 구성과 피치 추정

2.1 청각모델의 구성

본 연구에서는 청각의 특성을 구현하는 모델로서 기저막의 운동, 기저막과 유체압력과의 상호작용, 그리고 헤어셀 변환으로 이루어진 그림 1과 같은 Robinson 모델을 구성하였다.

* 正 會 員 : 延世大 電氣·컴퓨터工學科 博士課程
** 正 會 員 : (주) 비젼아이트 이사. 工博
*** 正 會 員 : 京東大 情報通信工學部 敎授·工博
§ 準 會 員 : 延世大 電氣·컴퓨터工學科 碩士課程
§§ 正 會 員 : 延世大 電氣·컴퓨터工學科 敎授·工博
接受日字 : 2000년 4월 10일
最終完了 : 2000년 5월 19일

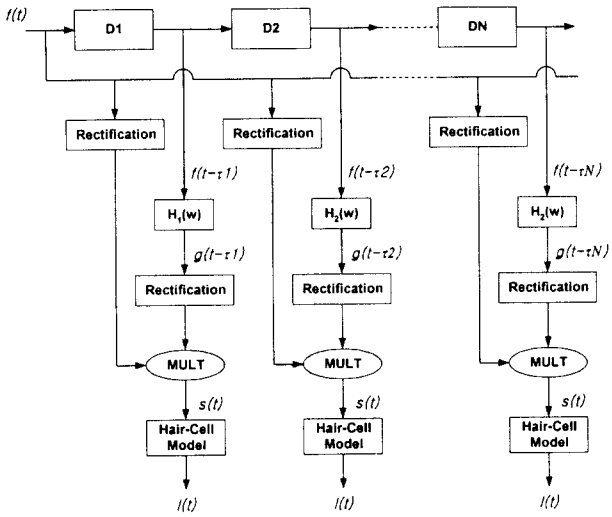

그림 1 Robinson 모델
Fig. 1 The Robinson model

2.1.1 기저막의 운동

시간의 함수로서 기저막 위를 진행하는 신호는 그 주파수 성분과 일치하는 특정주파수의 지점에 도달하면 최대 공진을 일으킨다. 따라서 기저막의 한 지점(요소)의 모델링은 지연단과 대역통과여파기의 연결로 구현될 수 있다[3].

각 요소의 지연시간은 최초로 입력신호를 받아들이는 스테이피즈 연결점에서부터 그 요소까지의 각 지연단의 지연시간의 선형적인 합과 같고 이 관계는 식 1과 같이 나타내어진다.

$$\text{n번째 요소에서의 지연시간 } \tau_n = \sum_{i=0}^{n} \Delta_i \qquad (1)$$

여기서 Δ_i는 각 요소의 지연시간이다.

그리고 각 대역통과여파기의 중심주파수는 지연시간의 역수로 정해지며 이를 그 요소의 특성주파수로 정한다. 여기서 w_{n0}는 각 대역통과여파기의 중심주파수이다

$$\text{특성주파수 } CF = \frac{1}{\tau_n} = \frac{w_{n0}}{2\pi} \qquad (2)$$

2.1.2 유체압력과의 상호작용

기저막상의 한 요소에 위치한 헤어셀의 섬모는 2가지 운동 즉, 섬모 주위의 임파액의 압력변화와 기저막의 공진운동에 의해 움직이게 된다. 1985년 Delaware 대학의 Yegnanaraynanan은 귀의 피치 인지원리를 설명하기 위해 이 두 가지 운동이 헤어셀에 미치는 상호작용을 반파정류와 곱셈으로 설명하였다[3]. 임파액의 운동은 스테이피즈로부터 전달된 기계적 진동 $f(t)$와 동일하며 기저막의 공진운동의 결과는 지연단과 대역통과 여파기를 거친 $g(t-\tau_n)$이다. 그림 1에서 $f(t)$는 임파액운동 즉, 스테이피즈로부터 전달된 기계적 진동을 나타내며 $g(t)$는 기저막의 공진운동의 결과를 나타낸다.

2.2 청각 모델을 이용한 피치 추출

2.2.1 Robinson 모델을 이용한 피치 추출 알고리즘

그림 2는 이렇게 구성된 청각모델에 200Hz의 정현파를 입력으로 사용하였을 때의 출력이다. 100Hz에서 5kHz까지의 특성주파수를 갖는 99개의 신경의 출력을 51.2msec 동안 나타내었다. 그림 2에서 신경출력도의 지연축을 보면 서로 인접한 여러 개의 신경들의 첨두값이 몇 개의 지점에 밀집되어 있음을 알 수 있다. 그리고 이 지점은 50 표본 지연으로 식 2에 따라 고려해 보면 200Hz와 그 고조파(harmonics)가 있는 위치임을 알 수 있다. 또한 시간축을 보면 정현파의 주기가 되는 5ms 간격으로 출력이 크게 나타난다. 즉, 정현파의 주기가 나타나는 시간 값에서 200Hz 주파수 부분에 해당하는 지연 부분이 가장 큰 출력을 나타낸다.

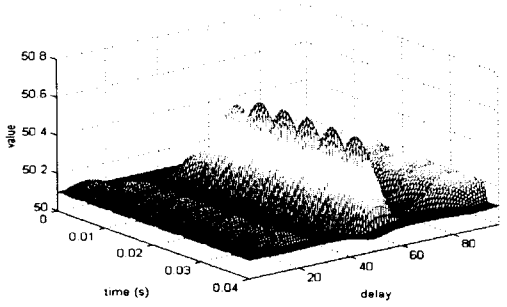

그림 2 200Hz 정현파에 대한 신경출력도
Fig. 2 Neurogram of sinusoidal wave

음성의 경우, 시간 축에서 볼 때 크기는 다르지만 일정한 주기성이 나타나지만, 여러 주파수의 신호가 합쳐진 것이므로 200Hz 정현파와는 달리 그 주기를 바로 알기는 쉽지 않을 것이다. 따라서, 그 주기성을 나타내기 위해서 각 지연에 대해서 자기상관을 취하는 방법을 사용하였다.

대부분의 지연값마다 일정의 주기값이 나오기는 하겠지만, 어떤 지연값에서는 주기성이 나오지 않는 경우도 존재하고, 또한 모든 지연값에서 자기상관 연산을 하여 주기성을 구하는 것은 연산량에도 문제가 될 수 있다. 따라서 본 논문에서는 에너지 값이 크게 나오는 지연값의 주변 값에서만 자기상관을 취하는 방법을 제안한다. 정현파의 경우 에너지가 크게 나타나는 부분의 지연값에 해당하는 주파수가 청각 모델에 입력된 신호의 주파수와 일치하는데서 착안한 것이다.

2.2.2 Cochannel 상에서의 피치 추정

최근에 들어서 중첩된 음향 신호를 각각의 음원 성분으로 분리하는 기술이 광범위하게 연구되어지고 있다[4-5]. 이러한 기술은 실생활에 적용되는 여러 신호처리 분야, 이를테면 cross-talk 음성인식, 잡음 하에서의 음성 인식 등에 있어서 매우 중요하다. 두 화자의 음성이 동시에 들리는 cochannel에서 각 화자의 음성을 분리하는 것은 이러한 경우 중에서도 특이한 경우이다.

이러한 cochannel 상에서의 음성분리를 위해서는 각각의 음원에 대한 피치 추정은 화자 분리의 기본요소로 사용된다. 피치 주파수 간격의 값들만 취하여 음성을 분리하는

표본 주파수법(sampling frequency method)을 근간으로 하는 경우는 피치의 추정이 잘못되면 합성신호의 음질이 차이가 발생하게 된다[4]. 따라서, cochannel 상에서 피치를 분리하는 연구가 진행되어야 했고, 최근에는 중첩된 신호에서 동시에 두 사람의 피치를 구하는 알고리즘도 연구되어지고 있다[6]. 본 연구의 목적은 cochannel 상에서 동시에 두 사람의 피치를 구하는 것이 아니라 가장 가능성이 높은 한 화자의 피치를 추정하는 것이다.

Cochannel 음성에 자기상관법을 사용하면 섞인 두 음성의 SIR에 따라 한 화자의 피치도 추정이 불가능할 경우도 발생한다. 본 연구에서는 cochannel 음성에 대하여 청각 모델을 사용하여 2장에서 제안한 알고리즘에 따라 피치를 추정하여 보고, 각 화자의 음성에서 추정한 피치와 비교하여 알고리즘의 타당성을 살펴볼 것이다.

3. 실험 및 결과

3.1 실험 환경

본 연구의 목적은 유성음에 대한 피치의 추정에 있으므로, 유성 단음에 대한 것으로 국한한다. 음성 데이터는 성인 화자 5인으로 하여금 /아/, /에/, /이/, /오/, /우/ 5개의 기본 모음을 발음하게 하여 이를 녹음하였다. 각 모음의 데이터는 10kHz의 주파수로 샘플링, 16bit로 양자화하여 얻었다. Cochannel에서의 피치 추출에 대한 실험을 위해서는 피치가 다른 두 화자의 음성의 데이터를 합하여 사용하였다.

3.2 청각모델을 통한 피치 추정

그림 3은 두 화자의 음성 중 실험에 쓰인 /아/ 발음의 정상상태 500개의 표본을 나타낸 것이다. 그림 4는 이 음성에 대한 청각모델의 출력을 나타낸 것이다. 그림 5는 이를 지연축에서 본 모습을 나타낸 것이다.

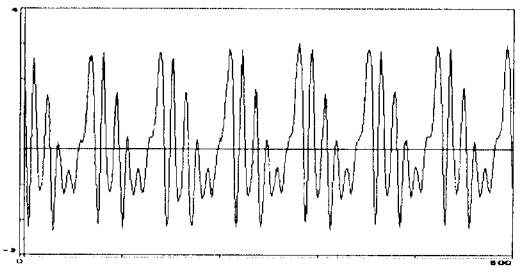

그림 3 화자 1의 /아/ 음성
Fig. 3 Speech /a/ of speaker 1

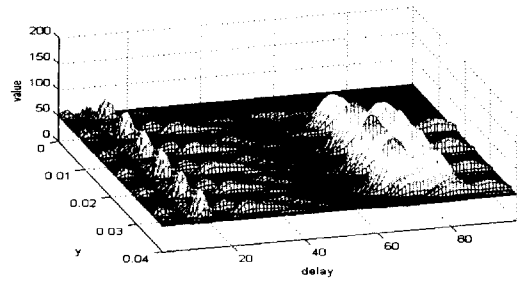

그림 4 /아/에 대한 청각 모델 출력
Fig. 4 Neurogram of speech /a/

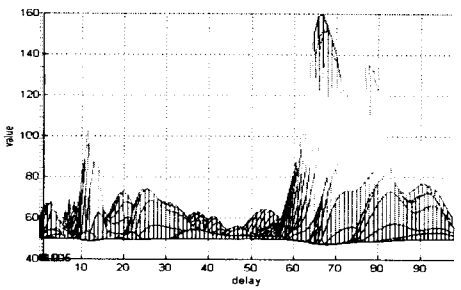

그림 5 /아/의 지연 축 neurogram
Fig. 5 Neurogram of /a/ in delay axis

지연축에서 본 그림 5를 보면, 가장 높은 첨두값의 지연값은 67이다. 이에 해당하는 주파수로 전환하면 식 2에 의해서 147.1Hz에 해당하는 값을 나타낸다. 이 것을 시간의 값으로 전환하면 6.7ms가 되며, 이는 자기상관으로 구한 값, 7.1ms와 거의 같은 값임을 알 수 있다. 이와 같은 현상을 이용하여 피치를 추정하려는 노력은 1987년에도 있어왔다[7].

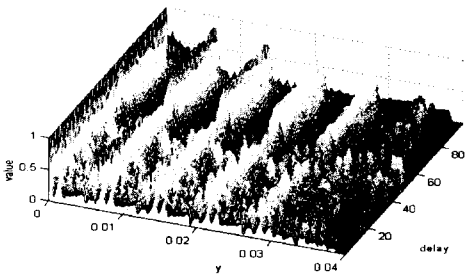

그림 6 /아/의 neurogram에 대한 자기상관 함수
Fig. 6 Autocorrelation function of the neurogram for /a/

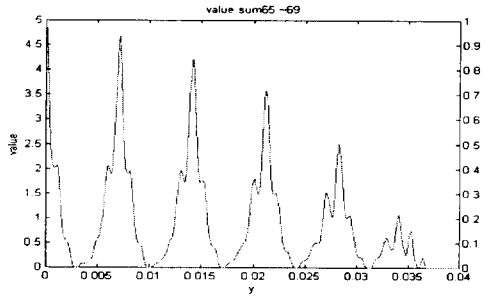

그림 7 첨두 주변값들의 합(/아/)
Fig. 7 Sum of peak neighbours(/a/)

그림 6은 본 연구에서 제안한 방법대로 /아/의 청각 발화 출력을 지연값마다 자기상관을 취하여 정규화한 것을 나타낸 것이다. 이 경우 주기성이 명확히 나타나지만 모든 지연값에 대해서 이렇게 자기상관기(autocorrelator)를 사용한다면 계산량이 많아지게 되고, 고주파 부분에서 자기상관기를 사용하면 많은 첨두값이 발생하여 오히려 주기성을 불명확하게 한다. 따라서 본 연구에서는 청각모델의 출력값들 중에서 최고값이 나타나는 부분의 지연값 주변의 값들에서만 자기상관기를 사용할 것을 제안하였다.

청각 모델 출력을 나타낸 그림 5에서 최고값을 가지는 부분의 지연값은 67이다. 이 지연값 주변의 자기상관 값을 합하여 나타낸 것이 그림 7이다. 주기성이 보다 명확하게 나타나는 것을 볼 수 있다. 다음 표 1은 자기상관법과 제안한 방법을 사용하여 추정한 피치 값을 나타난다.

표 1 피치 추정 결과

Table 1 Results of the pitch estimation

방 법	아	이	우	에	오
자기상관법(ms)	7.1	7.63	7.33	7.7	7.3
청각모델링(ms)	6.8	7.5	7.2	7.5	7.1

3.3 Cochannel 상에서의 피치 추출 실험

Cochannel 상에서의 피치 검출의 결과는 다음과 같다. 청각 모델의 피치 추정에 사용한 두 화자의 /아/와 /이/ 발음을 합한 신호를 제안된 방식에 적용하였다. 이때 두 음의 에너지 비 즉, SIR은 0.97 정도로 두 음이 거의 같은 에너지를 갖도록 하였다.

그림 8은 자기상관법을 적용한 것을 나타낸다. 그림 9-10은 청각 모델의 출력 및 제안한 방식으로 구한 피치 값을 나타낸 그림이다.

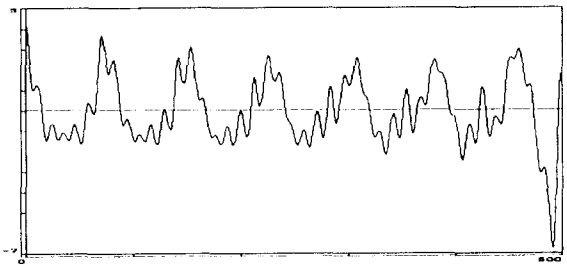

그림 8 두 화자의 유성음 /아/와 /이/의 자기상관 함수

Fig. 8 Autocorrelation function of two speaker vowels /a/ and /i/

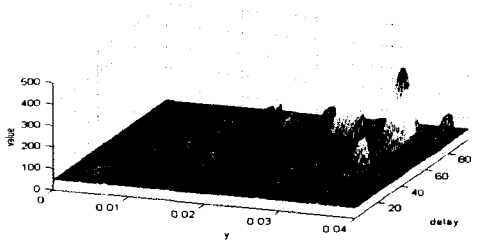

그림 9 Cochannel 음성의 신경출력도

Fig. 9 Neurogram of cochannel speech

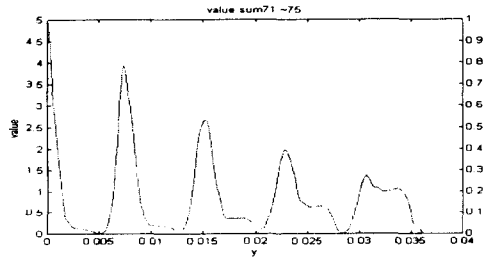

그림 10 첨두 주변값들의 합(/아/+/이/)

Fig. 10 Sum of peak neighbour(/a/+/i/)

그림 8에서 첫 번째 첨두가 첫 화자의 피치이고 그 옆의 조금 작은 첨두가 두 번째 화자의 피치라고 하기에는 그 첨두 간격이 일정치 않다. 그림 11은 합하기 전의 /이/ 발음신호에 피치 추출 알고리즘을 적용한 것을 나타낸 것이다.

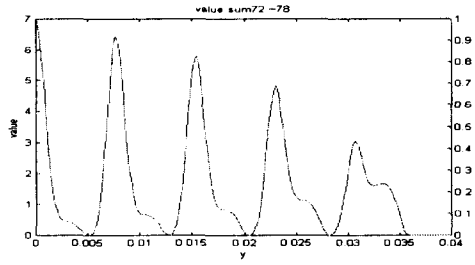

그림 11 첨두 주변값들의 합(/이/)

Fig. 11 Sum of peak neighbours(/i/)

Cochannel에서 구한 그림 10의 값과 그림 11, 즉 두 화자 음성 중 /이/ 발음의 피치가 거의 일치함을 알 수 있다. 다른 cochannel 음성들에 대한 실험에서도 비슷한 결과를 보여 한 화자의 피치는 거의 정확히 추출할 수 있었다.

참 고 문 헌

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

[2] Obaidat MS, Brodzik A, Sadoun B, "A performance evaluation study of four wavelet algorithms for the pitch period estimation of speech signals", Information Sciences, Vol.112 No.1-4 , pp. 213-221 , 1998, 12.

[3] G. Yegnanarayanan, "A new model of hearing and its performance in pitch perception", ph. D. thesis, Delaware Univ., 1985.

[4] R. J. McAulay and T. F. Quatieri, "Speech Analysis-Synthesis based on A Sinusoidal Representation", IEEE Trans. Acoust., Speech, Signal Processing, vol. 34, pp. 744-754, Aug. 1986.

[5] David P.Morgan, E. Bryan George, Leonard T. Lee, and Steven M. Kay, "Cochannel Speaker Separation by Harmonic Enhancement and Suppression", IEEE Trans. Speech, Audio Processing, Vol. 5, No. 5, pp. 407-424, 1997, 9.

[6] Matti Karjalainen and Tero Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis", in Proc. Int. Conf. Acoust., Speech, Signal Processing, vol. 2, pp. 929-932, 1999, 3

[7] Lee, Jae Hyuk, "The Study on the Speech Recognition using Auditory Model", M. S. thesis, Yonsei Univ., 1987.