2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

음색변경을 위한 피치시점 검출에 관한 연구

박형빈, 배명진
숭실대학교 정보통신공학과

On a Detection of Pitch Point for Voice Color Conversion

HyungBin Park, MyungJin Bae
Dept. of Telecom. Engr., Soongsil Univ. Seoul 156-743, Korea
hbpark@assp.soongsil.ac.kr

Abstract

음성신호처리분야에서 피치시점 검출은 음성 합성시에 여기원의 특성을 나타내어 음질의 자연성을 결정한다. 이에 본 논문에서는 음색 변경시에 운율조절에 필요한 피치시점 검출법을 제안한다. 제안한 방법은 시간영역에서 직접 처리하기 때문에 피치동기분석이 용이하고 다른 영역으로의 변환과정이 불필요하다. 또한 기존의 피치시점검출 방법에서는 결정논리를 실험적인 문턱값이나 무게치를 적용하여 처리하는 반면에 제안한 방법은 분석구간별로 얻어지는 주기적인 성문특성을 적용하여서 정확한 피치시점을 검출할 수 있었다.

Ⅰ. 서 론

음성신호처리분야에서 피치시점을 정확히 검출하는 것은 아주 중요하다. 피치시점을 정확히 검출할 수 있다면 음성분석시 피치동기된 분석을 할 수 있고, 인식시에는 성문의 영향이 제거된 정확한 성도 파라메터를 얻을 수 있으므로 인식의 정확도를 높일 수 있다. 또한 합성시에는 여기원의 위상특성을 파악할 수 있으므로 개성이 강조된 합성음을 얻을 수 있다[4].

Strube는 성대 폐쇄의 순간이 위치하는 음성파형의 공분산 행렬(Covariance matrix)에 대한 log행렬식(Determinant)의 측정을 제안했다. 이 방법은 모든 모음신호들에 대해서 적용할 수 없다. 다시 말해서 실제 어떤 모음들에서는 GCI(Glottal Closure Instant)를 결정하기가 매우 어렵기 때문이며 또한 처리시간이 많이 소요된다. Wong등은 음성파형의 M-점 창에서 p-pole 전체 선형 예측 에러 시컨스를 해석하는 방법을 제안했다. 그러나 이 방법은 성문의 폐쇄된 위상이 매우 짧은 구간을 가진 고주파나 호흡(breathy) 음성인 경우에서는 정확한 폐쇄위상을 얻는 것이 어렵다. Veeneman은 EGG신호를 이용하는 방법을 제안하였다. 하지만 이 방법은 후두에 마이크로폰을 부착하여 직접적으로 성대의 움직임을 측정하여야 하며 역 필터링(Inverse Filtering)과정이 필요하다[4][5].

일반적으로 음성신호에서 성도의 전달함수와 음원의 각 부분을 독립적으로 가정한다면 음성출력을 음원이 여과기를 통과하여 나오는 산호로 볼 수 있다. 따라서 음성신호에서 음원의 특징을 측정하기 위해서 역 필터링과정을 수행하기도 한다. 또한 성문파의 주기동안 시불변특성(Time Invariant)을 갖는다는 전제하에 처리되어졌다[6]. 하지만 이러한 방법은 여성화자와 같이 높은 피치를 갖는 음성신호에는 적합하지 않다. 따라서 이러한 역 필터링의 제한을 극복하기 위해서 성도필터와 성문파형의 파라메터를 동시에 측정하기 위한 방법을 연구하였다. 그러므로 성도전달함수는 일정한 폐위상 동안에(Closed Phase) 제한되어 측정된다. 이러한 단점에도 불구하고 역 필터링 과정은 많은 음성처리분야에서 널리 쓰이고 있다[5].

본 논문에서는 화자의 음색변경시 운율조절에 필요한 피치시점 검출법을 제안하였다. 먼저 음성신호의 발성모델에 근거하여 예측 계수(Predictor Coefficients)를 갖는 시변 자동회귀(Auto-Regressive)모델을 적용하여 역 필터링 과정을 수행하였다. 그런 다음 음성신호에서 한 피치구간에서의 주기적인 성문특성을 적용하여서 피치시점검출 과정을 수행하였다.

Ⅱ. 음성생성모델 및 선형예측분석

음성신호를 음성생성모델 측면에서 고려한다면 무성음의 경우에는 불규칙잡음생성기가 그 생성원이므로 주기성은 나타나지 않지만 주로 3kHz근방에서 공진 봉우리를 갖기 때문에 유성음에 비해서 평균 영교차율이 크

다[1][2]. 반면 유성음의 경우에는 성대의 진동과 성도의 공명 현상 때문에 시간영역에서 진폭이 크고 준주기적 성질을 나타낸다. 즉 유성음은 성문펄스가 그 생성원이며 성대의 진동에 따른 성도의 영향이 강조되어 나타난다. 다음 그림 2-1은 선형음성생성모델을 나타낸 것이다[6].

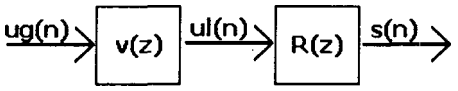

그림 2-1. 선형음성생성모델의 블록 다이어그램

선형음성생성모델에서 성도모델 v(z)는 전극모델(All-pole model)은 다음 식(2-1)과 같다.

$$V(z) = [\ 1 + \sum_{i=0}^{p} c_i\, z^{-i}\ ]^{-1},\ k=0,1,...,p-1 \qquad (2\text{-}1)$$

성대 펄스열 ug(n)은 성도 전달함수 v(z)로 필터링되고 구강의 결과는 ul(n)이다. 입술에서의 방사임피던스 R(z)는 다음 식(2-2)와 같다.

$$R(z) = 1 - z^{-1} \qquad (2\text{-}2)$$

또한 미분화된 성문파형, q(n)은 다음 식(2-3)과 같이 나타낼 수 있다.

$$q(n) = u_G * r(n) \qquad (2\text{-}3)$$

여기서 *는 길쌈(Convolution)과정을 나타낸다.

미분화된 성문파형, q(n)을 고려하면 음성생성모델은 다음 식(2-4)와 같이 나타낼 수 있다.

$$s(n) = \sum_{i=1}^{p} a_i(n)s(n-i) + q(n) \qquad (2\text{-}4)$$

인근한 음성신호들은 높은 상관관계를 가지고 있다고 가정한다면 다음 식(2-5)와 같이 표현할 수 있다.

$$y_n \cong \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + \cdots + \alpha_p y_{n-p} \qquad (2\text{-}5)$$

위 수식에서 음성신호의 표본화된 값(y_n)은 상수 α가 곱해진 과거의 p표본들에 의해서 예측할 수 있다는 가정을 보여주고 있다. 이 최소자승오차를 갖는 상수들을 선형예측계수라 하고 이 계수를 구하는 방법을 선형예측분석방법이라고 한다[1]. 또한 음성파형의 표본값을 위와 같이 선형예측분석하는 것은 다음 식(2-6)처럼 자동회귀모델을 갖는 전극 시스템이라고 가정한다.

$$y_n + \sum_{i=1}^{p} \alpha_i\, y_{n-i} = x_n \qquad (2\text{-}6)$$

위에서 설명한 선형예측계수를 구하는 방법에는 크게 자기상관(Auto-Correlation)법과 공분산(Covariance)방법이 있다[1][2]. 음성신호의 표본화된 값(y_n)이 충분히 길고 안정한 상태일때는 두 방법의 결과는 거의 차이가 없지만 표본화된 입력의 길이가 짧고 불안정한 상태라면 공분산 방법은 짧은 시간적인 변화에 좋은 응답을 주나 항상 안정된 결과를 주지 못한다.

Ⅲ. 음색변경을 위한 피치시점 검출법

음색변경을 위해서는 화자간의 성도 특성 외에도 운율정보의 변환도 고려되어야 한다. 또한 운율정보 변환시 먼저 피치검출과정이 수행되어져야만 한다. 하지만 분석프레임간 평균피치정보는 음성신호에서 음소변화 특성등을 잘 표현하기 어렵다. 따라서 정확한 피치시점을 검출할 수 있다면 피치동기 분석할 수 있고 운율정보의 변환이 용이하다.

음성생성모델의 관점에서 음성신호는 앞서 언급한 바와 같이 여기신호가 성도특성을 나타내는 필터를 통과함으로써 발생되는 신호로 볼 수 있다. 일반적으로 음색변환을 하기 위해서는 음성생성모델에 근거해서 사람의 성도특성을 전극필터로 가정하고 선형예측분석을 적용한다. 다음 식(3-1)과 같이 선형예측계수로 표현되는 필터에 역으로 통과시킴으로써 여기신호 특성을 잘 나타내는 잔여신호(Residual signal)를 얻을 수 있다.

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \qquad (3\text{-}1)$$

결과적으로 음성의 여기원으로 볼 수 있는 주기적인 펄스열들을 얻을 수 있다. 단일 입력 임펄스에 의한 출력의 범위에서 잔여 신호는 예측오차인 임의의 작은 랜덤한 파형처럼 보인다. 그러나 새로운 입력 임펄스가 더해질 때 예측오차는 증가된다. 이러한 입력 임펄스열은 성문의 떨림에 의해서 발생하는 피치 펄스열과 같다. 다음 그림 3-1은 선형예측분석을 통해서 얻은 예측신호의 예를 나타낸 것이다.

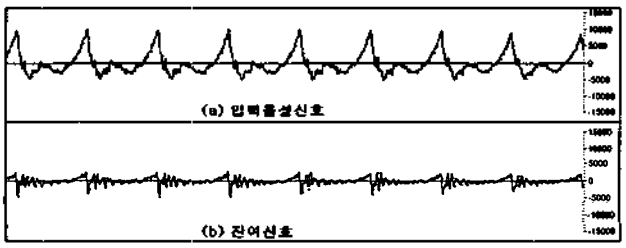

그림 3-1. 선형예측계수를 적용해서 역필터링 한 예

따라서 큰 예측오차는 입력 피치 펄스와 동기된 잔여신호에서 나타나고 유성음의 경우에서도 입력 또는 음원이 특정 피치 주기의 임펄스 열로 나타난다. 특히 여성 음성의 경우와 같이 짧은 피치구간의 경우에는 상기의 임펄스 응답이 다음 입력 펄스가 발생할 때까지 무시될 수 있을 정도로 충분히 작아지지 않기 때문에 분석 오류를 발생시킬 수 있다. 또한 떨리는 음이나 파열 폐쇄음과 같이 빠른 전이구간에서도 분석의 어려움이 있다.

본 논문에서는 단구간(short-term) 분석 잔여신호열을 가지고 피치동기된 분석을 통해서 피치시점 검출법을 제안하였다. 제안한 방법은 다음과 같이 크게 분석과정, 예측과정, 피치시점 검출과정으로 나누어진다.

Ⅲ-1. 분석 과정(Analysis)

선형예측계수를 적용한 역필터링과정에 의해서 얻어진 잔여신호 e(n)에서 예측피치신호 위치열 $\tilde{p}(n)$을 구하기 위해서 다음 식(3-2)과 같이 먼저 음(negative)의 값만을 고려한다. 이것은 양의 값에 비해서 예측 에러 값이 크기 때문이다. 또한 해당프레임에서 음의 평균 진폭값 이상으로 처리함으로써 부성문(Sub-Glottal)의 성분을 제거하였다.

$$\tilde{p}(n) = Position\ of\ e(n) < Ave.\ of\ the\ Neg.\ amp. \quad (3\text{-}2)$$

여기에서

$$Ave.\ of\ the\ Neg.\ amp. = -(\sum_{n=0}^{FrameSize} e(n))/Frame\ Size$$

Ⅲ-2. 예측 과정(Prediction)

분석 과정에서 구한 예측피치신호 위치열 $\tilde{p}(n)$이 실제로 음성의 여기원, 피치펄스 위치열 $p(n)$이라면 최소 피치구간 $\Delta\ Pitch_{min}$이상을 갖는 주기적인 특성을 나타낼 것이다. 하지만 분석과정에서 부성문의 성분을 제대로 제거하지 못했다면 피치주기내에 존재할 수 있기 때문에 다음 식(3-3)을 통해서 피치펄스 위치열 $p(n)$의 예측 에러를 최소화하였다.

$$\begin{aligned} p(n) &= |\ \tilde{p}(1) - \tilde{p}(2)\ | < \Delta\ Pitch_{min} \\ &= |\ \tilde{p}(2) - \tilde{p}(3)\ | < \Delta\ Pitch_{min} \\ &= |\ \tilde{p}(3) - \tilde{p}(4)\ | < \Delta\ Pitch_{min} \\ &\quad \vdots \\ &= |\ \tilde{p}(n-1) - \tilde{p}(n)| < \Delta\ Pitch_{min} \end{aligned} \quad (3\text{-}3)$$

$\Delta\ Pitch_{min}$는 한 피치주기내에서 존재할 수 있는 최소의 피치주기를 말한다. 일반적으로 음성신호에서의 피치범위는 2.5ms~25ms이므로 제안한 방법에서는 2.5ms를 최소피치주기로 하였다.

Ⅲ-3. 피치시점 검출 과정(Detection)

제안한 방법에서는 분석 및 예측과정을 통해서 얻은 피치펄스 위치열 $p(n)$을 통해서 피치시점을 검출하였다. 피치동기된 분석을 수행하였기 때문에 다음 식(3-4)처럼 원신호에서 해당프레임의 영교차 정보 $ZCI_{positive}(i)$ (+Positive ZCI)를 적용해서 정확한 피치시점을 검출할 수 있었다.

$$\begin{aligned} Pitch_{point}(m) &= ZCI_{positive}(i) < p(n) < ZCI_{positive}(i+1) \\ PITCH &= Pitch_{point}(m) - Pitch_{point}(m-1) \end{aligned} \quad (3\text{-}4)$$

다음 그림 3-2와 3-3은 제안한 방법을 적용한 경우 피치시점을 검출한 후 피치변화도를 나타낸 것이다.

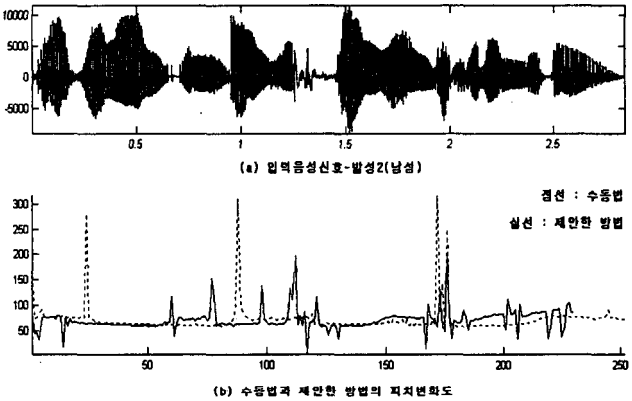

그림 3-2. 제안한 방법의 예(남성-발성2)

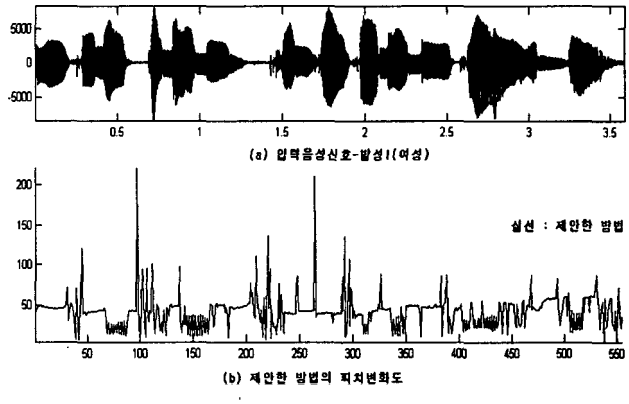

그림 3-3. 제안한 방법의 예(여성-발성1)

상기 그림을 보면 알 수 있듯이 남성의 경우에는 상당히 우수한 결과를 얻을 수 있었으나 여성의 경우에는 상대적으로 낮은 결과를 얻었다. 이것은 남성의 경우보다 여성의 경우가 피치주기가 짧고 앞서 언급한 바와 같이 빠른 전이구간에서 분석 오차로 인해서 약간의 예측에러가 발생하였다.

Ⅳ. 실험 및 결과

컴퓨터 시뮬레이션에 이용한 장비는 IBM-PC(P-Ⅱ) 시스템이며 여기에 음성신호를 입출력하기 위한 상용화된 16비트 AD/DA변환기를 인터페이스 하여 11kHz의 표본율로 데이터를 입력하였다. 각 시료에 대해 한 프레임의 길이를 320표본으로 하여 예측차수(p)만큼 프레임 오버랩과정을 수행하였다. 처리결과의 성능을 위해서 다음의 대표적인 문장들을 연령층이 다양한 남녀 5명 화자가 발성하여 시료로 사용하였다.

발성1: /인수네 꼬마는 천재소년을 좋아한다./
발성2: /여기는 음성통신 연구실입니다./
발성3: /예수님께서 천지창조의 교훈을 말씀하셨다./

제안한 방법을 구현하기 위해서 C-언어로 구현하여 수행하였다. 성능비교를 수행하기 위하여 수동으로 피치 시점을 구하였다. 다음 그림 4-1은 제안한 방법의 블록도를 나타낸 것이다.

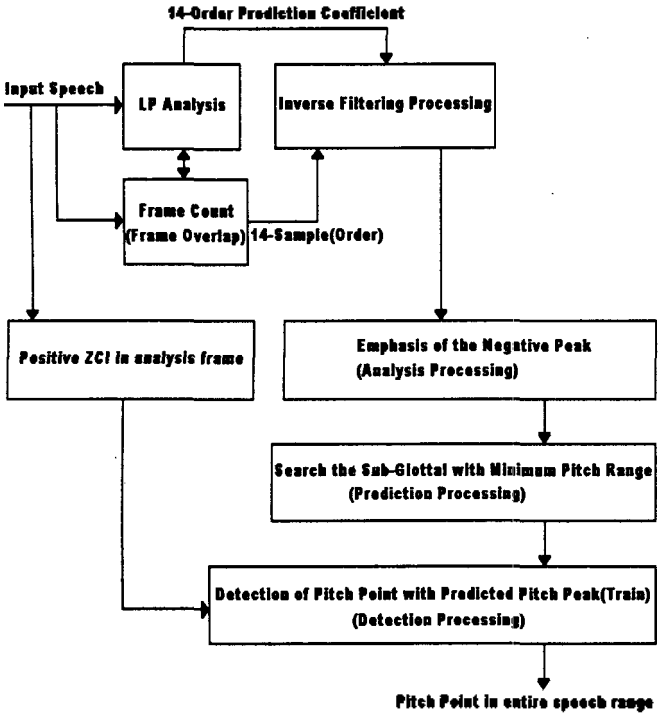

그림 4-1. 제안한 방법의 블록 다이어그램

Ⅴ. 결 론

음성신호처리분야에서 피치시점을 정확히 검출하는 것은 아주 중요하다. 피치시점을 정확히 검출할 수 있다면 음성 분석시 피치동기된 분석을 할 수 있고, 합성시에는 여기원의 위상특성을 파악할 수 있으므로 개성이 강조된 합성음을 얻을 수 있다[4].

본 논문에서는 음색변경시에 운율조절에 필요한 피치시점 검출법을 제안하였다. 제안한 방법은 시간영역에서 직접 처리하기 때문에 피치동기분석이 용이하고 다른 영역으로의 변환과정이 불필요하다. 또한 기존의 피치시점검출 방법에서는 결정논리를 실험적인 문턱값이나 무게치를 적용하여 처리하는 반면에 제안한 방법은 분석구간별로 얻어지는 주기적인 성문특성을 적용하여서 정확한 피치시점을 검출할 수 있었다.

하지만 피치주기가 짧고 음소의 변화가 빠르고 심한 전이구간을 가진 여성의 경우에는 남성의 경우에 비해 상대적으로 낮은 결과를 얻었다. 향후에는 음성의 분석구간을 사전에 결정하여서 안정구간이 아닌 경우에는 예측피치열을 강조함으로써 예측에러를 줄일 수 있는 세세한 후처리 과정에 대한 연구가 수행되어야 할 것이다.

Ⅵ. 참고문헌

[1] L. R. Rabiner and R. W. Schafer, Digital Proce-ssing of Speech Signal, Prentice Hall, 1978.

[2] J. D. Markel and A. H. Gray, jr., Linear Prediction of Speech Signals, Springer-Verlag, 1976.

[3] Hans Werner Strube, "Determination of the instant of glottal closure from the speech wave", J.Acoust.Soc.Am., Vol.56, No.5, pp.162-1629, November 1974.

[4] 이해군, 배명진, 임운천, "G-Peak 검출에 의한 음성신호의 피치시점검출", 제6회 신호처리합동학술대회, 제6권, 1호, pp.58-61, 1993.

[5] E. L. Riegelsberger, A. K. Krishnamurthy, "GLOTTAL SOURCE ESTIMATION:METHODS OF APPLYING THE LF-MODEL TO INVERSE FILTERING", ICASSP, Vol.2, pp.542-545, 1993.

[6] Xiao-Lin Tian, Matti Karjalainen, "Estimation of Glottal Source waveforms From Speech Signal Using Orthogonal Search Method", ICSPAT, Vol.1, pp.131-136. 1994.