자동 피치시점 검출에 의한 음성신호의 지속시간 조절법에 관한 연구

박 원, 박형빈, 배명진
숭실대학교 정보통신공학과

On a Duration Control Method of Speech Waveform by an Automatic Pitch Point Detection

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

Abstract

일반적으로 고음질 음성합성을 하기 위해서는 합성음의 지속 시간을 변경하여 줌으로써 운율을 조절하는 기법이 필요하다. 이에 먼저 고음질용 음성부호화법을 선정하여야 하고 정확한 피치와 피치시점검출을 통해서 음원분류가 되어야한다. 본 논문에서는 제안한 자동 피치시점 검출을 적용해서 운율조절에 필요한 지속시간 조절법을 제안하고자 한다. 제안한 방법은 시간영역에서 직접 처리하기 때문에 피치동기분석이 용이하고 다른 영역으로의 변환과정이 불필요하다. 결과적으로 파형부호화법을 적용하고 제안한 자동 피치서점 검출에 의한 지속시간 조절법을 적용하였을 때 비교적 우수한 결과를 얻을 수 있었다.

Ⅰ. 서 론

최근 다양해진 음성합성에서는 고음질의 합성음을 요구하고 있다. 고음질 합성을 위해서는 합성음의 지속시간을 변경하여 줌으로써 운율을 조절하는 기법이 필요하다. 이러한 기법을 이용하면 명료한 발음속도 조절을 통해 음성에 의한 검색이 가능하며 음성을 압축하는데도 크게 도움이 된다. 또한 언어장애인의 발음교정이나 어학 학습을 위해서는 음성의 지속시간변경이 필수적이라 할 수 있다.

기존에 사용되고 있는 지속 시간 변경법으로는 신호원부호화를 이용한 방법과 파형부호화를 이용한 방법이 있다. 신호원부호화를 이용한 방법은 음성 생성 모델상의 음원 여기구간을 원하는 만큼 지속시킴으로써 지속시간을 변경한다. 이들 알고리즘은 전송율을 낮출 수 있기 때문에 채널 대역폭이나 메모리 효율적인 방법이지만 분석시에 각 성분을 분리하고 다시 그 정보를 이용해서 합성하기 때문에 분석시의 오차와 합성시의 오차가 합해져서 합성음질은 자연성이나 명료성이 크게 떨어진다. 또한 피치시점검출의 정확도는 합성음질의 자연성에 크게 영향을 미친다.

이러한 기존의 지속 시간 변경법에 있어서의 문제점은 고음질용 부호화법을 선정하여야 하고 음원분류가 정확해야 한다는 것이다. 그리고 정확한 피치와 피치시점검출이 이루어져야 한다는 것이다. 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].

일반적으로 음성신호에서 성도의 전달함수와 음원의 각 부분을 독립적으로 가정한다면 음성출력을 음원이 여파기를 통과하여 나오는 신호로 볼 수 있다. 따라서 음성신호에서 음원의 특징을 측정하기 위해서 역 필터링과정을 수행하기도 한다[5].

본 논문에서는 제안한 자동 피치시점 검출을 적용해서 운율조절에 필요한 지속시간 조절법을 제안하였다. 제안한 방법은 음성신호의 발성모델에 근거하여 예측계수(Predictor Coefficients)를 갖는 시변 자동회귀

(Auto-Regressive)모델을 적용하여 역 필터링 과정을 수행하였다. 그런 다음 음성신호에서 한 피치구간에서의 주기적인 성문특성을 적용하여서 피치시점을 검출하였다. 이렇게 검출된 피치시점을 사용하여 원하는 변경율 만큼 피치주기단위별로 삭제 또는 반복함으로써 지속시간을 변경 시키는 방법을 제안하였다.

Ⅱ. LP 분석에 의한 자동 피치시점 검출법

일반적으로 인근한 음성신호들은 높은 상관관계를 가지고 있다고 가정한다면 다음 식(2-1)과 같이 표현할 수 있다.

$$y_n \cong a_1 y_{n-1} + a_2 y_{n-2} + \cdots + a_p y_{n-p} \qquad (2\text{-}1)$$

위 수식에서 음성신호의 표본화된 값(y_n)은 상수 a가 곱해진 과거의 p표본들에 의해서 예측할 수 있다는 가정을 보여주고 있다. 이 최소자승오차를 갖는 상수들을 선형예측계수라 하고 이 계수를 구하는 방법을 LP(Linear Prediction, 선형예측) 분석방법이라고 한다[1]. 또한 음성파형의 표본값을 위와 같이 선형예측분석하는 것은 다음 식(2-2)처럼 자동회귀모델을 갖는 전극시스템이라고 가정한다.

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

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.

또한 운율변환을 위한 지속시간 변경시 먼저 피치시점검출과정이 수행되어져야만 한다. 하지만 분석프레임간 평균피치정보는 음성신호에서 음소변화 특성등을 잘 표현하기 어렵다. 따라서 정확한 피치시점을 검출할 수 있다면 피치동기 분석할 수 있고 운율정보의 변환이 용이하다.

일반적으로 다음 식(2-3)과 같이 선형예측계수로 표현되는 필터에 역으로 통과시킴으로써 여기신호 특성을 잘 나타내는 잔여신호(Residual signal)를 얻을 수 있다.

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \qquad (2\text{-}3)$$

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 2-1은 선형예측분석을 통해서 얻은 예측신호의 예를 나타낸 것이다.

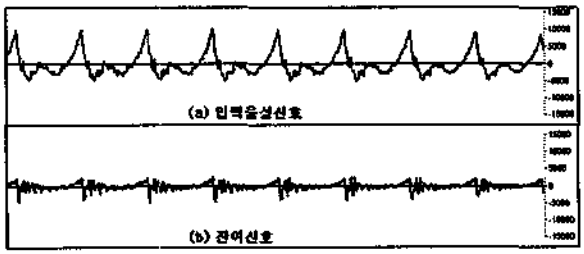

그림 2-1. 선형예측계수를 적용해서 역필터링 한 예

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

Ⅱ- 1. 분석 과정(Analysis)

선형예측계수를 적용한 역필터링과정에 의해서 얻어진 잔여신호 e(n)에서 예측피치신호 위치열 $\tilde{p}(n)$을 구하기 위해서 다음 식(2-4)와 같이 먼저 음(negative)의 값만을 고려한다. 이것은 양의 값에 비해서 예측 에러값이 크기 때문이다. 또한 해당프레임에서 음의 평균진폭값 이상으로 처리함으로써 부성문(Sub-Glottal)의 성분을 제거하였다.

$$\tilde{p}(n) = Position\ of\ e(n) < Ave.\ of\ the\ Neg.\ amp. \qquad (2\text{-}4)$$

여기에서

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

Ⅱ- 2. 예측 과정(Prediction)

분석 과정에서 구한 예측피치신호 위치열 $\tilde{p}(n)$이 실제로 음성의 여기원, 피치펄스 위치열 $p(n)$이라면 최소피치구간 $\triangle Pitch_{min}$이상을 갖는 주기적인 특성을 나타낼 것이다. 하지만 분석과정에서 부성문의 성분을 제대로 제거하지 못했다면 피치주기내에 존재할 수 있기 때문에 다음 식(2-5)를 통해서 피치펄스 위치열 $p(n)$의 예측 에러를 최소화하였다.

$$
\begin{aligned}
p(n) &= |\ \tilde{p}(1)-\tilde{p}(2)\ | < \Delta\ Pitch_{min} \\
&= |\ \tilde{p}(2)-\tilde{p}(3)\ | < \Delta\ Pitch_{min} \\
&= |\ \tilde{p}(3)-\tilde{p}(4)\ | < \Delta\ Pitch_{min} \\
&\vdots \\
&= |\ \tilde{p}(n-1)-\tilde{p}(n)\ | < \Delta\ Pitch_{min}
\end{aligned} \qquad (2\text{-}5)
$$

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

Ⅱ- 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(2-6)처럼 원신호에서 해당프레임의 영교차 정보 $ZCI_{positive}(i)$ (+Positive ZCI)를 적용해서 정확한 피치시점을 검출할 수 있었다.

$$
\begin{aligned}
Pitch_{point}(m) &= ZCI_{positive}(i) < p(n) < ZCI_{positive}(i+1) \\
PITCH &= Pitch_{point}(m) - Pitch_{point}(m-1)
\end{aligned} \qquad (2\text{-}6)
$$

다음 그림 2-2는 제안한 방법을 적용한 경우 피치시점을 검출한 후 피치변화도의 예를 나타낸 것이다.

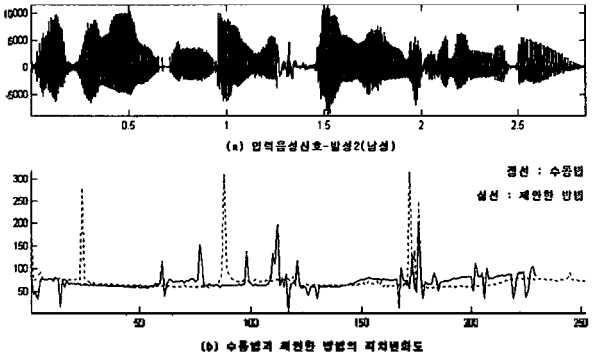

그림 2-2. 제안한 피치시점 검출방법의 예(남성-발성2)

Ⅲ. 지속시간 변경법

일반적으로 고음질용 파형부호화를 이용한 방법은 한 피치주기의 파형을 반복시키는 방법으로 피치를 결정할 수 없는 폐쇄음이나 파열음 등은 별도로 처리해야 하는 어려움이 따른다. 신호원부호화를 이용한 방법과 마찬가지로 피치검출의 정확도가 자연성에 지대한 영향을 준다. 피치의 시점은 피치의 반복에 중요인수로 작용한다.

본 논문에서는 앞서 언급된 제안한 자동 피치시점 검출을 통해서 예측된 피치시점위치열을 구한다. 검출된 피치시점위치열들이 준주기적인 형태를 통해서 유/무성음을 구별한다. 하지만 제안한 피치시점검출법은 성문파형에서의 G-Peak의 진폭특성과 주기적인 성문특성을 통해서 자동으로 피치시점 위치열을 얻기 때문에 그러한 특성이 반영이 되지않은 구간에서는 합성음의 명료성을 유지하기 위해서 지속시간 변경과정을 수행하지 않았다. 다음 식(3-1)은 본 논문에서의 지속시간 변경법에 적용한 변경율을 나타낸 것이다.

$$
D_{ratio} = \frac{P_{point}(N)}{D_{factor}} \qquad (3\text{-}1)
$$

여기에서 $Dratio$ 는 지속시간 변경율, $P_{point}(N)$ 는 입력신호에서 검출된 예측피치시점위치열의 전체 개수, $Dfactor$ 는 지속시간 변경시 피치주기단위별로 삭제 또는 반복하는 횟수를 나타낸다.

분석구간이 음소가 빠르게 변하는 천이구간이거나, 지속시간 변경율이 높은 경우에는 인근한 피치구간별로 에너지 레벨변화가 심하기 때문에 구간에서의 레벨을 평균화 시켜서 합성음의 음질 저하를 줄였다.

앞서 언급한 바와 같이 고음질 합성을 위해서 합성음의 지속 시간을 변경하여 줌으로써 운율을 조절하는 기법에서는 명료성이 유지되어야 한다. 본 논문에서 제안한 자동 피치시점 검출법과 지속시간 조절법은 음성신호에서 주기적인 성문파형의 진폭특성과 주기특성에 의해서 예측피치시점열을 구하였고 준주기적인 형태를 갖는 분석 프레임 구간에 대해서만 지속시간 변경과정을 수행하였다. 상대적으로 명료성이 유지되는 우수한 합성음을 얻을 수 있었다.

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 45ms으로 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./
발성4: /감사합니다./

제안한 방법을 구현하기 위해서 C-언어로 구현하여 수행하였다. 성능비교를 수행하기 위하여 수동으로 피치시점을 구하였다. 다음 그림 4-1은 제안한 자동 피치시점 검출방법의 블록도를 나타낸 것이다.

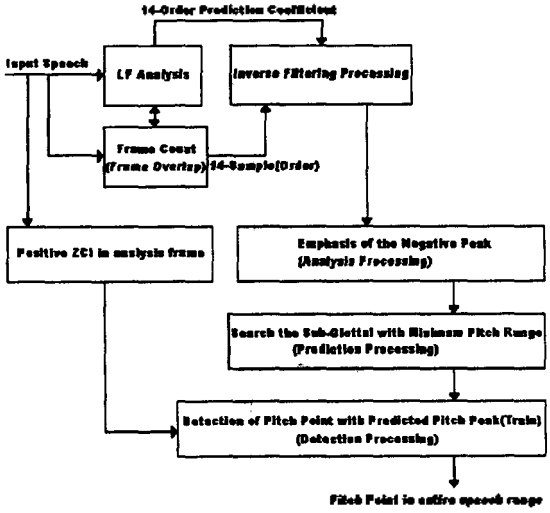

그림 4-1. 제안한 자동 피치시점 검출 방법의 블록 다이어그램

다음 그림 4-2와 4-3은 제안한 방법을 적용해서 음성신호의 지속시간을 변경한 예를 나타낸 것이다.

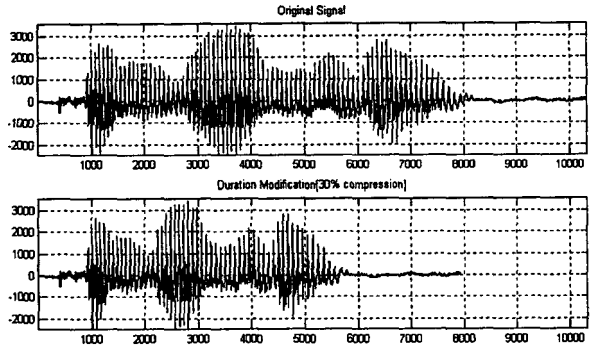

그림 4-2. 제안한 지속시간 변경의 예(1) - 발성4

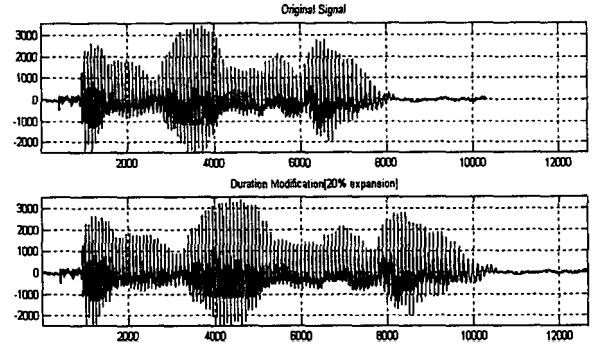

그림 4-3. 제안한 지속시간 변경의 예(2) - 발성4

V. 결 론

최근 다양해진 음성합성에서는 고음질의 합성음을 요구하고 있다. 고음질 합성을 위해서는 합성음의 지속시간을 변경하여 줌으로써 운율을 조절하는 기법이 필요하다. 이러한 기법을 이용하면 명료한 발음속도 조절을 통해 음성에 의한 검색이 가능하며 음성을 압축하는데도 크게 도움이 된다. 또한 언어장애인의 발음교정이나 어학 학습을 위해서는 음성의 지속시간변경이 필수적이라 할 수 있다.

또한 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 제안한 자동 피치시점 검출을 적용해서 운율조절에 필요한 지속시간 조절법을 제안하였다. 제안한 방법은 시간영역에서 직접 처리하기 때문에 피치동기분석이 용이하고 다른 영역으로의 변환과정이 불필요하다. 결과적으로 파형부호화법을 적용하고 제안한 자동 피치서점 검출에 의한 지속시간 조절법을 적용하였을 때 비교적 우수한 결과를 얻을 수 있었다.

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.