스펙트럼 보상된 고음질 합성용 피치 변경법

°문효정, 배재옥, 김용, 배명진
숭실대학교 정보통신공학과

On a Pitch Alteration Method Compensated with the Spectrum for High Quality Speech Synthesis

Hyojung MOON, Jaeok BAE, Yong KIM, Myungjin BAE
Dept. of Telecommunication, Soongsil University

* 본 논문은 1992년도 한국과학재단의 연구비 지원으로 이루어진 것임.
(과제번호: 92-21-00-06)

ABSTRACT

The waveform coding are concerned with simply preserving the wave shape of speech signal through a redundancy reduction process. In the case of speech synthesis, the waveform coding with high quality are mainly used to the synthesis by analysis. However, because the parameters of this coding are not classified as either excitation and vocal tract parameters, it is difficult to applying the waveform coding to the synthesis by rule. In this paper, we proposed a new pitch alteration method that can change the pitch period in waveform coding by using scaling the time-axis and compensating the spectrum. This is a time-frequency domain method that is preserved in the phase components of the waveform and that has a little spectrum distortion with 25% and less for 50% pitch change.

Ⅰ. 서론

음성 합성분야에서 합성단위에 따라서는 문장단위, 음절단위, 음소단위 등의 합성법으로 나눌 수 있다. 또한 합성방식에 따라서는 파형부호화법, 신호원부호화법, 혼성부호화법으로 분류할 수 있다[1-3]. 파형부호화법은 파형 자체의 잉여성분을 제거한 후에 부호화 하는 방법이며, PCM, ADPCM, ADM등이 제안되어져 있다. 이 부호화법은 인간의 개성과 감정을 대변해 주는 여기정보와 의사전달을 나타내는 성도의 여파기정보를 분리하지 않고 처리하기 때문에 음원을 변경시켜야 하는 음절단위나 음소단위의 합성기법으로는 바람직하지 못하다.

신호원부호화법은 음성발성모델에 따라 분석 시에 여기정보와 여파기정보를 분리시켜서 독립적으로 부호화 하는 방법으로 LPC, PARCOR, LSP, 포만트 등의 알고리듬이 제안되어져 있다. 신호원부호화법은 분석시에 성분을 분리하고, 다시 그 정보를 이용해서 합성하기 때문에 분석시의 오차와 합성시의 오차가 합해져서 합성음질은 자연성이나 명료성이 크게 떨어진다.

신호원부호화의 메모리 효율성과 파형부호화의 명료성 및 자연성을 적당히 유지하기 위해서 이 두 가지 부호화기법을 결합시킨 방법으로 혼성부호화법이 있는데 이들에는 MPLPC, RELP, VELP, CELP 등의 방법이 있다. 그렇지만 혼성부호화법에서 성도 여파기정보의 부호화에는 신호원부호화법을 적용하고, 성대 여기정보의 부호화에는 파형부호화법을 주로 적용하고 있다. 따라서 파형부호화법과 마찬가지로 여기정보를 변경시켜야하는 음절단위나 음소단위의 규칙 합성방식에는 적용하기가 어렵다.

최근 다양해진 음성서비스 분야에서는 고음질의 합성음을 요구하고 있다. 이러한 고음질 합성방식으로는 파형부호화법이 바람직하다. 그렇지만 파형부호화법을 사용하면 상기와 같이 메모리 규모가 방대하고 음원피치의 변경이 어렵다는 문제점이 발생한다. 그러나 부호화에 필요한 메모리 문제는 현재의 기술수준으로 충분히 극복이 가능하다. 나머지 문제의 해결 법으로는 파형부호화법을 규칙에 의한 합성에 적용되도록 음원피치를 변경시킬 수 있어야 한다.

Ⅱ. 기존의 피치 변경법

파형부호화법이나 혼성부호화법은 분석 후 합성을 하는 문장단위의 합성법에는 오랫동안 적용되었으나, 음원피치의 변경이 용이하지 못하기 때문에 음절 및 음소단위의 합성법으로는 사용되지 못하고 있는 실정이다. 가끔 단어

나 반음절단위로 파형부호화법이나 혼성부호화법이 사용되고 있으나, 같은 단어라도 연결되는 단어에 따라 다른 데이타를 사용하고 있는 실정이다. 따라서 파형부호화에 의해 우수한 음질로 규칙에 의한 합성을 수행하려면 피치변경이 필요하다.

지금까지 제안된 피치변경법은 그 처리영역에 따라 시간, 주파수, 시간-주파수 혼성처리법이 있다. 시간영역법에는 Multi-Pulse법, 피치반분법 등이 있다. Caspers와 Atal은 MPLPC에서 펄스사이에 영을 삽입하거나 삭제하는 방법을 제안했으나, MPLPC상의 펄스 열은 피치와 포먼트에 대한 상호연관을 갖고 있으므로 스펙트럼의 왜곡이 심하다[5]. Varga와 Fallside는 LPC계수를 이용한 피치연장법을 제안했으나, 이 방법은 피치주기를 줄이는 경우에는 단지 파형의 일부분을 소거하고 평활화하는 방법을 사용하고 있기 때문에 스펙트럼의 왜곡이 많이 나타난다[6]. 피치반분법은 임의로 변경하려는 피치주기의 2배 파형을 만든 후에 그 파형의 주기를 반분하는 피치 변경법이다[8]. 그러나 이 방법은 시간영역에서만 수행되기 때문에 스펙트럼 왜곡이 발생하여 명료성이 저하된다.

Quatieri와 McAulay는 주파수영역에서 위상을 보존하는 피치변경법을 제안하였는데 이것은 입력된 음성에 대해 진폭 및 위상스펙트럼을 추출하여 별도로 처리하는 방법이다. 진폭스펙트럼에 대해서는 두드러진 스펙트럼 봉우리들을 추출한 다음에 이것을 피치변경율(ρ) 만큼 인터폴레이션하여 진폭스펙트럼의 피치를 변경시킨다. 위상스펙트럼에 대해서는 시간영역에서 구한 피치 개시시간(pitch onset time)에 해당하는 위상을 제거하고 나서 피치가 변경되었을 때의 새로운 피치 개시시간의 위상을 더해줌으로서 새로운 위상을 구성하게 된다. 이러한 방법은 파형의 꼴을 그대로 유지하기 때문에 프레임단위로 분석처리하는 통상의 처리법에서 인접 프레임간의 연결이 아주 용이해 진다는 장점이 있다. 그렇지만, 피치변경시에 피치주기와는 별도로 피치의 개시시간을 공급해 주어야 하고, 또한 진폭 스펙트럼 상에서 두드러진 봉우리 위주로 고조파의 인터폴레이션을 수행하기 때문에 스펙트럼의 왜곡이 높아진다는 단점이 있다.

다른 주파수영역 피치변경법으로는 평탄화기법에 의해 포먼트와 기본주파수의 고조파를 분리하여 기본주파수를 선형적으로 스케일링하므로써 피치를 변경하는 방법이 있다. 이 방법은 스펙트럼특성을 원래의 음성스펙트럼으로 대신하는 방법이나 스펙트럼 상에서 고조파를 스케일링시킴으로서 창함수의 특성도 변경되어 시간영역에서 위상을 복원하기가 어렵게 된다[14].

시간-주파수 혼성법으로는 켑스트럼의 특징을 이용하여 켑스트럼값이 거의 영이 되는 부분에서 영값을 삽입하거나 삭제하므로써 피치를 변경하는 방법이 있다[7]. 그러나 이 방법 역시 위상의 보존이 어렵다는 문제점을 가지고 있다. Takagi와 Miyasaka가 제안한 시간-주파수혼성법은 시간영역에서 피치변경을 하였을 때에 나타나는 스펙트럼 왜곡을 스펙트럼영역상에서 LPC포락을 통해 수정하는 방법이다. 이 방법은 LPC스펙트럼 포락이 갖는 국침에 치중된 시스템 전달특성 때문에 모든 유성음을 만족하지 못한다는 한계성을 갖는다.

피치변경시에 포만트 스펙트럼이 왜곡되면 성도의 여기정보가 변경되므로 의사정보들 제대로 보존할 수 없고, 위상이 왜곡되면 인근 프레임간 진폭레벨의 변동이 커져서 음소간의 연결이 부자연스럽게 된다. 따라서 본 논문에서는 스펙트럼의 왜곡을 최소화하면서도 시간영역의 위상을 그대로 보존할 수 있는 새로운 시간-주파수 혼성영역 피치변경법을 제안하고자 한다. 먼저 시간영역에서 시간축 스케일링에 대해 설명한 다음 새로이 제안하는 피치변경법에 대하여 알아본다. 그리고는 우리가 제안한 피치변경법을 실제 음성에 적용해 본 결과에 대해 비교 및 검토하였다.

III. 유성음의 시간축변경에 따른 스펙트럼왜곡

유성음은 성대의 진동과 성도의 공명을 통해서 발생하는 것이기 때문에 진폭스펙트럼상에 고조파들이 나타나고, 이들의 포락은 성도의 공진주파수 봉우리를 이루게 된다. 4KHz 이하에서는 2~4개 정도의 공진봉우리가 관찰되는데 이들은 낮은 주파수로부터 제 1포먼트(F1), 제 2포먼트(F2) 등으로 부른다. 강한 유성음구간에서는 F1이 다른 포먼트들에 비해 약 10dB정도 높게 나타나며, 영-주파수에서 F1까지 에너지가 증가되는 봉우리 형태를 띠게 된다.

단순한 평균값 인터폴레이션을 통해 유성음의 피치주기를 r배로 늘리게되면 피치주기는 자연스럽게 늘어나지만 성도특성이 변경되어 발성의 명료도가 저하하게 된다. 실험에 사용한 표준 문장의 발성에 대해 평균값 인터폴레이션법으로 피치주기를 늘렸을 때에 원래 음성에 대한 스펙트럼왜곡을 측정하여 100분 율로 표 3-1에 나타내었다. 스펙트럼측정은 원래음성과 피치변경된 음성의 진폭스펙트럼을 해당 피치간격으로 스므딩하여 프레임 단위로 차이를 측정한 것이다. 피치변경율이 증가할수록 스펙트럼의 왜곡 율은 지수 함수적인 증가추세를 나타내고 있고, 동일한 피치변경율에서는 피치주기가 짧은 여성화자의 경우가 스펙트럼왜곡이 높았다.

또한 표본 삭제형 데시메이션을 통해 유성음의 피치주기를 r배로 줄이게 되면 피치주기는 자연스럽게 압축되지만 성도특성이 변경되어 발성의 명료도가 저하하게 된다. 실험에 사용한 표준 문장의 발성에 대해 삭제형 인터폴레이션법으로 피치주기를 늘렸을 때에 원래 음성에 대한 스펙트럼왜곡을 측정하여 100분 율로 표 3-2에 나타내었다. 이 경우에도 동일한 피치변경율에서는 피치주기가 짧은 여성화자의 경우가 스펙트럼왜곡이 높았고, 또한 피치 신장시보다는 스펙트럼 왜곡 율이 조금씩 높았다.

표 3-1. 피치주기를 신장시킨 경우의 스펙트럼왜곡율

발성자	피치주기 신장율				
	20%	40%	60%	80%	100%
남성	2.3	3.7	6.6	12.2	15.7
여성	3.0	4.7	7.6	13.3	16.7
평균	2.7	4.2	7.1	12.8	16.2

표 3-2. 피치주기를 압축시킨 경우의 스펙트럼왜곡율

발성자	피치주기 압축율				
	20%	40%	60%	80%	100%
남성	3.1	4.2	7.5	14.1	17.9
여성	3.2	4.9	7.9	14.4	17.3
평균	3.2	4.6	7.7	14.3	17.6

Ⅳ. 피치 변경법

유성음의 스펙트럼은 성도의 특성을 나타내는 포만트들과 시간영역에서 피치주기에 해당하는 기본주파수의 고조파들로 구성된다. 피치주기를 변경하기 전에 원래 음성의 스펙트럼포락을 추출하여 피치변경시의 스펙트럼과 교환하게 된다. 피치주기를 알고 있을 때 주파수상에서 피치의 역수인 기본주파수 Fo의 간격을 표본간격 Ko로 나타내면 다음과 같다:

$$Ko = \frac{size}{pitch} \qquad (4\text{-}1)$$

여기서 size는 창함수의 길이를 나타낸다. 이때 포만트 스펙트럼의 대수포락 H(K)는 대수 진폭스펙트럼을 다음 Lifter 함수에 통과시키므로써 구할 수 있다.

$$H(K) = \frac{1}{Ko} \sum_{L=1}^{Ko} S(K - L) \qquad (4\text{-}2)$$

$$(K = 0,1,\ldots, size/2-1)$$

유성음신호가 시간축 r율로 변경된 경우에는 피치주기가 r율로 변경되지만, 포만트 스펙트럼의 왜곡이 초래된다. 왜곡된 포만트 스펙트럼을 원래 음성에서 구한 포만트 스펙트럼으로 대체하면 왜곡을 줄일 수 있다. 먼저 시간축 변경된 신호 s'(rt)를 주파수영역으로 옮기고 대수 진폭스펙트럼 S'(k)를 계산한다. 대수 진폭스펙트럼은 변경된 기본주파수의 고조파구조 E'(k)와 성도특성의 포락 H'(k)이 더해진 구조를 이룬다. 다음의 리프트함수에 대수 진폭스펙트럼을 통과시키면 성도포락 스펙트럼 H'(k)이 얻어진다:

$$H'(K) = \frac{1}{Ko'} \sum_{L=1}^{Ko'} S'(K - L) \qquad (4\text{-}3)$$

$$(K = 0,1,\ldots, size/2-1)$$

여기서 $Ko' = Ko/r$ 이다. 이때 피치변경된 고조파구조 E'(k)는 다음과 같다:

$$E'(k) = S'(k) - H'(k), \quad k=0,1,\ldots,size-1 \qquad (4\text{-}5)$$

이제 피치주기가 변경되고 스펙트럼왜곡이 최소인 대수 진폭스펙트럼을 구성하면 다음과 같다:

$$S''(k) = H(k) + E'(k), \quad k=0,1,\ldots,size-1 \qquad (4\text{-}5)$$

이 대수 진폭스펙트럼 S''(k)를 지수함수에 통과시켜 진폭스펙트럼을 만들고, 피치변경된 위상스펙트럼을 사용하여 IFFT를 수행하면 피치변경된 파형이 얻어지게 된다.

상기의 피치변경 과정에는 시간-주파수-시간영역의 변환이 수행되어야 한다. 이러한 과정에서 창함수의 영향을 받게 되고 이로 인해 시간영역으로 변환된 파형의 위상이 어긋나거나 또는 스펙트럼의 왜곡이 발생할 수 있다. 창함수의 영향으로는 창함수의 모양과 길이에 따른 영향이 지배적인데, 신호 파형의 피치주기의 배수와 창함수길이를 일치시키면 그 영향을 최소화할 수 있다. 즉, 주어진 프레임에 대해 시간 축을 일정 율로 좁디 변경시켜 피치주기와 창함수의 길이가 일치되도록 하여 시간-주파수-시간변환을 수행하고 다시 시간 축을 일정 율로 다시 되돌림으로서 보상하는 방법이다.

Ⅴ. 피치검출법

신호원부호화법과는 달리 파형부호화법에서 피치를 변경하려면 사전에 그 발성자의 피치변화를 알고 있어야 한다. 이것은 발성자의 억양이나 감정의 변화가 중심된 피치를 기준으로 하여 피치의 상대적인 변화로 나타나기 때문이다. 특히 파형부호화에서는 발성자의 개성과 메시지 정보를 보존하여 음질의 명료성이 우수하다. 이 때문에 피치변경시에는 발성자가 주로 사용하는 피치주기를 기준으로 피치를 변경시킬 필요가 있다. 따라서 정확한 피치검출이 선행되어야 한다.

지금까지 제안된 피치검출법은 크게 시간영역법, 주파수영역법, 그리고 시간-주파수 혼성영역법으로 나눌 수 있다[1-3]. 본 논문에서는 피치검출법으로 시간영역에서의 면적비교법을 적용하였다. 그렇지만 합성을 위해 파형을 편집하는 경우에는 피치의 추출이 반드시 자동화될 필요는 없으며, 면적비교법[9]과 함께 눈으로 피치를 추출하는 반자동 법이나, 눈으로 찾는 수동 법으로 처리하여도 된다.

Ⅵ. 실험 및 결과

본 논문에서 제안한 방법을 시뮬레이션하기 위해 IBM-PC/486DX2-66에 마이크 입력이 가능한 16bit A/D변환기를 인터페이스하여 5명의 남성과 여성화자를 통해 다음 음성시료를 발성하게 하고 이를 8KHz의 표본화 율로

16비트 양자화하여 저장하였다.

발성 1: /인수네 꼬마는 천재소년을 좋아한다./

발성 2: /예수님께서 천지창조의 교훈을 말씀하셨다./

발성 3: /숭실대 정보통신과 음성통신연구팀이다./

발성 4: /공일이삼사오육칠팔구/

발성 5: /창공을 헤쳐나가는 인간의 도전은 끝이 없다./

그림 6-1은 상기의 피치 변경과정을 블럭도로 나타낸 것이다. 시뮬레이션에는 한 프레임의 길이를 256표본으로 사용하였다. 먼저 면적비교법을 사용하여 한 피치구간의 음성표본을 피치단위로 자른 다음에 각 피치 구간마다 합성억양에 비례된 피치주기를 결정한다. 그런 다음 현재 피치 한 구간을 반복시켜 한 프레임 길이의 파형을 만들고, 또한 원하는 율로 시간 축을 변경하여 피치변경된 파형을 만든다. 이 두 파형을 각각 주파수영역으로 변환한 다음에 피치변경된 성도스펙트럼 대신에 원래 음성의 성도스펙트럼을 사용하여 시간영역으로 변환하여 피치주기가 변경된 파형을 취하게 된다.

피치주기를 10%에서 50%까지 변경시켰을 때에 원래 스펙트럼에 비해 나타나는 스펙트럼의 왜곡 율을 측정하여 백분율로 표 6-1에 나타내었다. 스펙트럼의 기준은 피치변경되기 이전의 원래 음성의 스펙트럼이었다. 피치를 변경시키면 원래의 스펙트럼과 직접 비교할 수 없기 때문에 피치주기를 10%, 20%, 30%, 40%, 50%로 각각 신장시킨 다음에 91%, 83%, 77%, 71%, 67%로 각각 압축하여 원래의 음성스펙트럼과 고조파를 일치시킨 다음에 100분율 에너지왜곡을 측정하였다. 표 6-1을 살펴보면 피치주기를 50% 변경시켜도 스펙트럼 왜곡 율이 평균 2.5% 정도로 나타났다.

VII. 결 론

일반적으로 주파수영역 피치변경법은 스펙트럼의 왜곡은 적으나 위상의 왜곡이 심하고 시간영역 피치 변경법은 위상의 보존은 용이하나 스펙트럼 왜곡이 심하다는 문제점이 있다. 따라서 본 논문에서는 스펙트럼 왜곡을 최소화하면서도 시간영역에서의 위상을 보존할 수 있는 새로운 피치 변경법을 제안하였다. 즉, 원래 음성의 성도 포먼트 성분을 피치변경된 스펙트럼과 교환함으로서 피치주기를 변경하였다.

제안된 피치변경법은 시간영역에서 시간축 변경을 통해 피치를 변경하기 때문에 파형의 위상특성을 그대로 보존할 수 있으며, 또한 주파수영역에서 스펙트럼 보상을 수행하기 때문에 피치주기를 50% 변경하여도 스펙트럼 왜곡율은 평균 2.5% 정도로 우수하게 얻어졌다.

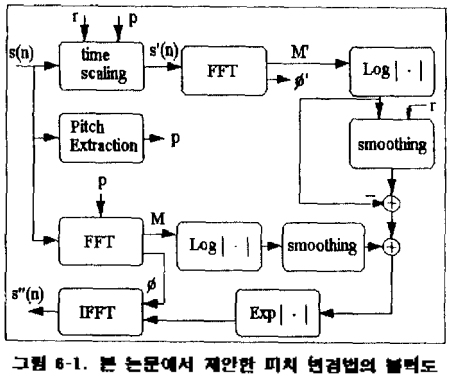

그림 6-1. 본 논문에서 제안한 피치 변경법의 블럭도

[REFERENCES]

[1] L.R. Rabiner & R.W. Schafer, *Digital Processing of speech Signals*, Prentice-Hall, 1978.

[2] M.R. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis," IEEE, Trans., Acoust. Speech, Signal Processing, Vol.ASSP-29, No.3, pp.374-390, June 1981.

[3] A. vargy and F. Fallside, "A Technique for Using Multipulse Linear Predictive Speech Synthesis in Text-to-speech Type System", IEEE signal processing, Vol.ASSP-35, No.4, pp.586-587, APRIL 1987.

[4] I.E. Caspers and B.S. Atal, "Changing Pitch and Duration in LPC Synthesised Speech using Multipulse Excitation", J. Acoust. Soc. Amer., Vol.73, No.1, pp.55, Spring, 1983.

[5] M.G. Stella and F.J. Charpentier, "Diphon Synthesis using Multipulse Coding and a Phase Vocoder", Proc. IEEE ICASSP'85, PP.740~744, 1985.

[6] 배명진, 이미숙, 이해균, 안수길, "캡스트럼 분석에 의한 음성 파형코딩의 피치변경에 관한 연구", 제 4 회 신호처리 합동 학술대회 논문집, 제 4 권 1호, pp.304~309, 1991년 9월.

[7] 강동규, 김윤재, 배명진, 안수길, "음성합성의 halving기법에 의한 파형 코딩의 피치 변경에 관한 연구", 한국음향학회 추계발표회(국제 음향학회 논문집), pp.107~111, 1990년 11월 10일.

[8] T. Takagi and E. Miyasaka, "A Speech Prosody Conversion System with a High Quality Speech Analysis - Synthesis Method," Proc. EUROSPEECH'93, pp.995~998, September 1993.

[9] 배명진, "위상보상된 고조파 스케일링에 의한 음성합성용 피치변경법", 한국음향학회, 한국음향학회지, 제 13권 6호, pp.91-97, 1994년 12월.

[10] 배명진, 안수길, "변적 비교법을 이용한 음성신호의 고속 피치 추출", 전자공학회지, 제22권, 2호, pp.13~17, 1985년 3월.

표 6-1. 피치변경에 따른 스펙트럼 왜곡율

변경율	여성화자	남성화자	평균(%)
10%	1.1	0.9	1.0
20%	1.4	1.3	1.4
30%	1.8	1.5	1.7
40%	2.1	1.9	2.0
50%	2.7	2.3	2.5