음성 파형분절의 지수함수 스무딩 기법에 관한 연구

(*) 박 찬 수 (*) 김 형 태 (*) 배 명 진 (**) 안 수 길
(*) 호서대학교 전자공학과 (**) 서울대학교 전자공학과

The Study on the Exponential Smoothing Method of the Concatenation Parts in the Speech Waveform

(*) Chansou PARK (*) Hungtae KIM (*) Myungjin BAE (**) Souguil ANN
(*) Hoseo University (**) Seoul University

Abstract

In a text-to-speech system, sound units (phonemes, words, or phrases, etc.) can be concatenated together to produce required utterances. The quality of the resulting speech is dependent on factors including the phonological/prosodic contour, the quality of basic concatenation units, and how well the units join together.

Thus although the quality of each basic sound unit is high, if occur the discontinuity in the concatenation part then the quality of synthesis speech is decrease. To solve this problem, a smoothing operation should be carried out in concatenation parts. But a major problem is that, as yet, no method of parameter smoothing is available for joining the segment together.

Thus in this paper, we proposed a new algorithm that smoothing the unnatural discontinuous parts which can be occurred in speech waveform editing. This algorithm used the exponential smoothing method.

I. 서 론

음성신호의 분석, 인식 및 합성기법에 대한 연구는 국외는 물론 국내에서도 많은 관심을 가지고 계속되고 있다. 일반적 합성기법으로는 문장단위의 합성, 음절단위의 합성, 음소단위의 합성으로 나누어 연구되고 있고, 최근에는 음소 단위로 합성하면서도 자연성과 명료성을 높이려는 연구가 국내외적으로 많이 제안되고 있는 중이다. 음성 시스템에서 필요한 발성을 만들어내기 위해 음소나 단어, 구와 같은 음성단위는 필요에 따라서 함께 연결될 수 있다. 이 때, 결과 음성의 음질(자연성과 명료성)은 음성학적/운율적인 변화에 포함된 인자들과 기본적인 연결단위의 음질, 그리고 단위가 얼마나 잘 결합되는가에 의해 좌우된다. 이들 인자들은 모두 상호작용을 하고 고음질 합성음을 얻기 위해서는 이 모든 것이 만족되어야 한다. 그러나 서로 다른 특성을 가지고 있는 파형을 자연스럽게 연결하는 것은 매우 어려운 일이다[1,2,3,4,5].

음성신호의 합성에 대한 코딩기법에는 소스코딩, 파형코딩, 혼성코딩법이 있다. 이들 중에서 파형코딩법은 음질이 우수하기 때문에 분석에 의한 합성코딩법으로 많이 적용되고 있지만, 음원과 성도의 특성을 분리하지 않고 피치의 잉여분만을 제거한 후에 파형자체를 코딩하기 때문에 규칙에 의한 합성코딩법으로 사용하기에는 어려움이 있고, 또한 소스코딩법에 비해 메모리 양이 많이 소요된다[6].

그렇지만 최근 메모리 반도체 제조기술이 발달함에 따라 메모리 양의 제한성을 해결해 주었고 다양해진 음성서비스 분야에서는 우수한 합성음질의 서비스를 요구하고 있기 때문에 파형코딩에 의한 합성법의 제안이 현실적으로 요구되고 있다. 파형코딩법으로 음성을 합성할 때는 파형자체의 음원과 성도의 성분을 분리하지 않고 피치주기를 변경시켜야 한다[6,7].

그러나 피치주기를 늘릴 경우에나 줄일 경우에는 주기의 끝에서 불연속성이 발생할 수 있다. 즉, 피치주기를 변경시킬 때는 성도의 특성이 영으로 수렴하지 않을 경우가 있다. 음성파형에서 이러한 불연속성은 주파수 영역에서 무한대의 주파수 특성을 나타내므로 음성의 명료성이

저하된다. 지금까지는 이러한 불연속성을 제거하기 위해서 선형적인 스므딩 기법을 주로 사용하여 왔다[6]. 이러한 방법 역시 파형 연결이 자연스럽지 못하여 스펙트럼의 왜곡을 초래한다.

따라서 본 논문에서는 음성파형 분절의 연결부분을 스므딩하기 위해 지수함수를 사용한 자연스러운 스므딩을 통해 스펙트럼상에서 발생하는 왜곡(distortion)을 줄이는 새로운 알고리즘을 제안하고자 한다.

Ⅱ. 음성신호의 피치 분절 방법

유성음의 피치구간 사이에는 성도(vocal tract)의 공명 특성을 나타내며, 이것은 안정시스템인 성도의 특성을 나타내기 때문에 다음 피치때 까지는 파형의 진폭이 점차 감쇄하는 모양이 된다. 따라서, 유성음의 다음 피치가 발생하기 전에 파형진폭은 거의 영에 근접하게 된다. 여기서 피치주기를 늘리려면 음성파형의 피치가 끝나는 부분에 늘리고자 하는 만큼의 영을 삽입하면 된다. 또한 피치주기를 파형상에서 줄이려면 일정 파형을 제거하는 방법을 사용하기도 한다[7].

시간영역상에서 직접 음성파형의 피치를 잘라내어, 잘려진 두 피치를 서로 결합하면 두 피치 사이의 연결부분에는 그림 1의 (b)와 같이 불연속점이 나타나게 된다. 이러한 결과로 인하여 음성 합성시에 명료성이 저하되어 음질이 낮아지게 된다. 그림 1의 (a)는 유성음 파형이며 (b)는 유성음의 피치를 줄이기 위해 (a)파형을 임의로 잘라 붙인 파형이다. 이 경우 음성의 피치는 쉽게 줄였지만 두 파형 사이의 연결부분이 자연스럽지 못하다.

잘라붙인 두 파형 사이의 연결부분을 선형적으로 대치하는 방법으로는 선형스므딩(linear smoothing) 기법이 있다. 이 방법은 두 피치 사이의 불연속점을 자연스럽게 연결하기 위한 방법의 하나로 다음 식에 의해 스므딩 한다 :

$$LS(n) = S(n-N) + \{ S(n)-S(n-N) \}(1- \frac{n}{N}) \qquad (2-1)$$

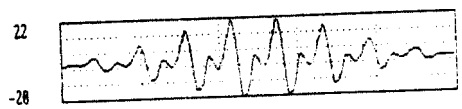
여기서, S(n)은 스므딩 구간의 초기값이고 N은 스므딩하고자 하는 구간이며, S(n-N)은 최종값이 된다. 그림 1의 (c)는 피치 결합부분에 (2-1)식을 적용하여 선형 스므딩한 음성파형을 나타낸다. 결합한 두 피치 사이가 선형적으로 연결되어 있음을 알 수 있다. 그렇지만 이 방법역시 그림에서 볼 수 있는 것처럼 파형의 시작부분과 끝부분에는 불연속적인 부분이 나타나게 된다.

Ⅲ. 지수함수 스므딩 기법

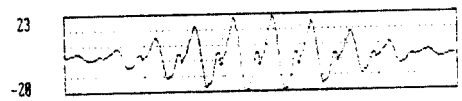
음성신호의 피치를 변경할 때, 파형 편집시 연결부분에 나타나는 불연속적인 부분을 스므딩하기 위해서 본 논문에서는 지수함수 기법을 다음식과 같이 제안한다 :

$$ES(n) = S(n-N) + \{ S(n)-S(n-N) \} e^{-\frac{n \times K}{N}} \qquad (3-1)$$

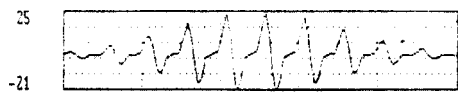
여기서 S(n)은 스므딩 구간의 초기값이고 N은 스므딩하고자 하는 구간이다. 또한 K는 스므딩 시정수 이고, S(n-N)은 스므딩 구간의 최종값이다. 따라서 두 피치 사이의 파형을 식 (3-1)을 적용하여 스므딩하면 그림 1의 (d)와 같이 곡선을 이루면서 자연스럽게 연결된다. 따라서 선형 스므딩에서 보다 분절된 부분이 원래의 파형 모양과 유사하게 스므딩되어 나타남을 알 수 있다.



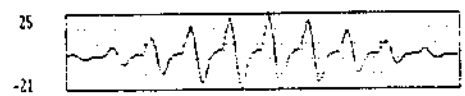
(a) 음성파형



(b) 피치를 스므딩 없이 임의로 잘라붙인 음성파형



(c) 피치를 임의로 잘라 선형 스므딩하여 붙인 음성파형



(d) 피치 연결부분에 지수함수 스므딩 기법을 사용하여 결합시킨 음성파형

그림 1. 음성파형 자체를 분절하는 방법들

음성파형의 피치 변화시, 두 피치사이의 연결 부분을 스므딩 하는 방법들과 본 논문에서 제안한 지수함수 스므딩 방법을 비교해 보았다. 우선, 각 방법에 대하여 결합된 피치들을 부분 반복시킨 후 각각에 대해 로그 스펙트럼을 구한다. 이 스펙트럼에 대하여 다시 역변환 하여 큐프런시(qufrency) 상에서 다음 식에 의해 왜곡을 측정하여 비교하였다 :

$$MCF = \frac{\sigma xy}{(\sigma x\ \sigma y)^{1/2}} \quad 100 \quad (\%) \qquad (3\text{-}2)$$

Ⅳ. 실험 및 결과

이상의 과정을 처리하기 위하여 마이크가 장착된 A/D 변환기를 IBM PC/AT에 인터페이스 시키고, 음성시료 1로는 24세의 남성화자가 연속 발성한 "인수네 고미는 연재소년을 좋아한다" 와 음성시료 2로는 28세의 남성화자가 연속 발성한 "호서대 전자공학과 음성신호처리 연구단이다"를 8KHz의 샘플링 주파수로 양자화하여 저장하였다.

본 논문에서 제안한 방법의 처리과정은 다음과 같다. 우선 세가지 방법의 피치 변환 방법을 비교하기 위하여 음성파형을 한 피치주기 단위로 편집을 하는 과정에서 불연속적인 부분이 발생하도록 논으로 피치를 살렸다. 첫번째 파형은 잘린 부분을 그대로 스므딩 없이 반복 연결하여 음성파형을 편집하였고, 두번째 파형은 식 (2-1)을 적용하여 다섯 샘플을 선형적으로 이어 부분 반복하여 음성파형을 생성하였다. 마지막으로 본 논문에서 제안한 방법인 지수함수 스므딩 기법을 적용하여 각 피치와 피치 사이를 식 (3-1)에 통과시켜 자연스럽게 연결되록 변형하였다.

위 세가지 음성파형에 대해 로그 스펙트럼을 구한 후 이를 다시 역변환 하여 큐프런시 상에서 식 (3-2)를 적용, 두 파형에 대한 왜곡(distortion)을 구하여 피치 연결기법들을 비교 평가 하였다.

그림 3은 발성 2에 대한 실험 결과이다. (a)는 음성음 파형을 나타내고 (b), (c), (d)는 (a)파형에 대한 피치를 8샘플 만큼 줄인 파형을 나타낸다. 그림 3의 (b)는 잘라낸 피치 부분을 스므딩 없이 그대로 붙여 연결한 파형이며, (c)는 연결부분을 선형 스므딩한 결과 파형이다. (d)는 본 논문에서 제안한 지수함수 스므딩 기법으로 처리한 결과이다. 그림 3의 스므딩 구간은 5샘플로 결정하여 처리한 결과이다. 또한 각각의 스므딩 기법을 비교하기 위하여 식 (3-2)를 적용, 왜곡을 측정하여 비교하였다. 이 결과 본 논문에서 제안한 지수함수 스므딩 기법이 다른 방법들에 비해 우수한 것으로 나타났다.

그림 4는 그림 3의 과정과 같은 방법으로 처리한 결과이며, 여기서는 스므딩 구간을 10샘플로 하여 처리하였다. 이 결과에서도 지수함수 스므딩 기법이 우수함을 입증하였다.

Ⅴ. 결론

음성 시스템에서 필요한 발성을 만들어내기 위해 음소나 단어, 구와 같은 음성단위는 필요에 따라서 함께 연결될 수 있다. 이 때, 결과 음성의 음질(자연성과 명료성)은 음성학적/운율적인 변화에 포함된 인자들과 기본적인 연결 단위의 음질, 그리고 단위가 얼마나 잘 결합되는가에 의해 좌우된다.

파형코딩을 이용한 음성합성은 자연성이나 명료성이 뛰어나며, 피치의 주기를 변경시킨다면 더욱 우수한 음질을 보장할 수 있다. 그러나 피치를 늘리고 줄이는데 있어서 파형의 연결부분에서 불연속적인 부분이 발생한다.

시간영역에서의 이러한 불연속은 주파수 영역에서 무한대의 주파수 성분으로 나타나므로 명료성을 최대한 보장하기 위해서는 두 파형의 연결부분에서 스므딩이 수행되어야 한다.

본 논문에서는 파형 분절의 연결부분을 스므딩하는 방법으로 지수함수 스므딩 기법을 적용하여 두 파형을 자연스럽게 연결해 줌으로써 파형 편집시 발생할 수 있는 음성파형분절의 연결부분을 스므딩하는 알고리즘을 제안하였다.

이 알고리즘을 적용한 결과 파형편집시 연결부분에서의 불연속성을 제거할 수 있었고 합성음에 대한 자연성과 명료성 저하의 문제를 해결할 수 있었다.

[REFERENCE]

[1] L.R.Rabiner & R.W.Schafer, *Digital Processing Processing of Speech Signal*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1987.

[2] S.D.Stearns & R.A.David, *Speech Signal Processing*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.

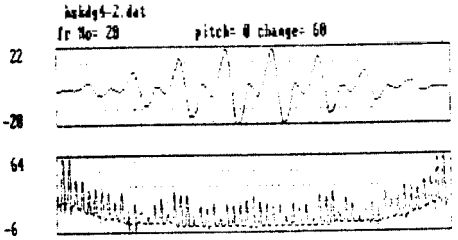
[3] P.E.Papamichalis, *Practical Speech Processing*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.

[4] Douglas O'Shaughnessy, *Speech Communication (human and machine)*, Addision-Wesley Publishing Company, 1987.

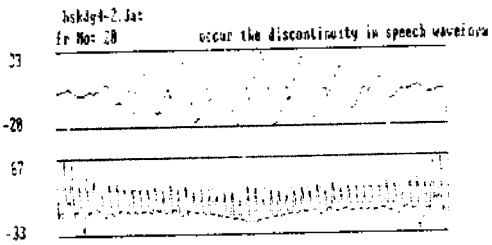
[5] J.D.Markel,A.H.Gray, *Linear Prediction of Speech*, Springer-Verlag Berlin Heidelberg ,New York, 1976.

[6] ANDRE VARGA & FRANK FALLSIDE," A Technique for Using Multipulse Linear Predictive Speech Synthesis in Text-to-Speech Type System", ASSP-35., NO.4, APPIL 1987.

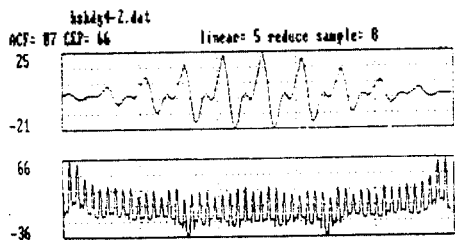
[7] 강동규,김윤제,배명진,안수길," On a pitch change of the waveform coding by the halving method for for speech waveform", 국제 음향학술발표회 문집 pp.107-111, 1990.



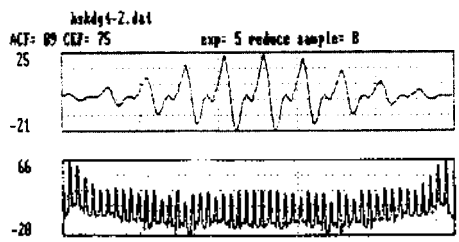
(a) 음성파형



(b) 피치를 스므딩 없이 임의로 갈라붙인 음성파형

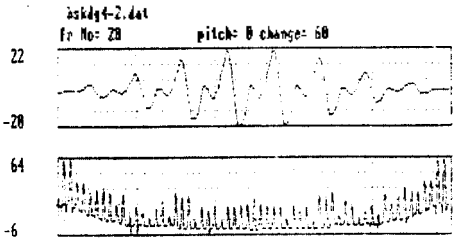


(c) 피치를 임의로 갈라 선형 스므딩하여 붙인 음성파형

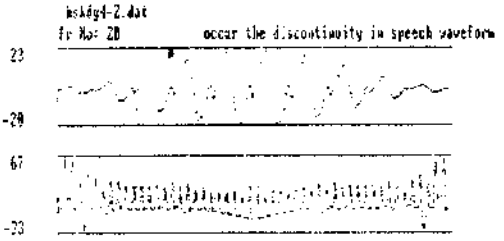


(d) 피치 연결부분에 지수함수 스므딩 기법을 사용하여 결합시킨 음성파형

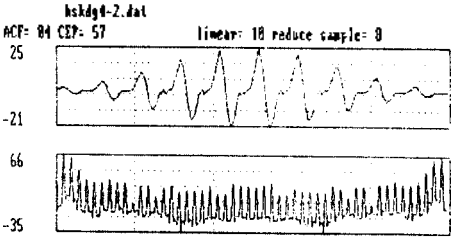
그림 3. 음성파형 자체를 분절하는 방법들(5샘플 스므딩)



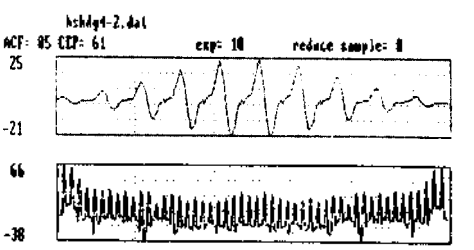
(a) 음성파형



(b) 피치를 스므딩 없이 임의로 갈라붙인 음성파형



(c) 피치를 임의로 갈라 선형 스므딩하여 붙인 음성파형



(d) 피치 연결부분에 지수함수 스므딩 기법을 사용하여 결합시킨 음성파형

그림 4. 음성파형 자체를 분절하는 방법들(10샘플 스므딩)