Analysis-By-Synthesis/OverLap-Add(ABS/OLA) Sinusoidal Model 을 이용한 음성변환과 연결음성합성

구자형, 최무열, 김형순
부산대학교 전자공학과

Speech Modification and Concatenative Speech Synthesis by using Analysis-By-Synthesis/OverLap-Add(ABS/OLA) Sinusoidal Model

Ja Hyoung Koo, Mu Yeol Choi, Hyung Soon Kim
Dept. of Electronics Eng., Pusan National University
E-mail : {jhkoo, shallom, kimhs}@hyowon.cc.pusan.ac.kr

요약

Sinusoidal Model 은 음성 신호처리의 넓은 분야에 적용되고 있는 방법으로 고음질의 합성음을 생성해 낼 수 있고, 조작이 용이하다는 장점을 가지고 있다. 본 논문에서는 Analysis-by-Synthesis/OverLap-Add(ABS/OLA) Sinusoidal Model 이라는 방법을 이용하여 시간축 변환과 음성 변환을 수행하였다. 특히 본 논문에서는 음질향상을 위하여, 시간축 변환시에는 정적인 구간과 변화하는 구간을 구별하여 서로 다른 시간축 변환비를 이용하였고, 기존의 LPC 방법에 비해 스펙트럼 포락선을 보다 잘 추정하는 Improved Cepstrum 을 이용하여 음정변환에 적용하였다. 또 서로 다른 문맥에서 얻어진 음성 단위들을 결합할 때 생기는 위상차이를 극복하기 위하여, 기본주파수 성분이 일치하도록 시간축을 이동하여 합성하였다. 실험결과 본 논문에서 적용한 방법들을 통해 기존 방식에 비해 개선된 음질을 얻을 수 있었다.

I. 개요

Sinusoidal Model 은 각기 다른 주파수, 진폭, 그리고 위상을 가지는 정현파들의 합으로 음성신호를 모델링하는 방법으로써, 이 방법은 고음질의 합성음을 나타내면서도 음의 고저 조절이 자유롭고, 또한 인접 음들 사이의 연결부분에서의 부자연스러움을 효율적으로 극복할 수 있는 장점이 있다[1]. 음성신호로부터 각 정현파들의 파라메터들을 추출하기 위한 방법에는 이산 Fourier 변환을 이용한 스펙트럼 영역에서의 피크추출 방법과 Analysis-by-Synthesis(ABS) 방법이 있는데, 후자의 경우가 더 우수한 것으로 알려져 있다[3]. 본 논문에서는 ABS 방법을 이용하여 정현파들의 파라메터를 추출하고, 프레임들을 가산중첩(Overlap-Add)함에 의해서 음성을 합성하는 방법을 이용하였다.

음성신호의 음정 변환시에는 스펙트럼 포락선을 유지하면서 기본주파수만이 바뀌도록 해야한다. 이렇게 하기 위해서는 스펙트럼 포락선 정보를 이용해야 하는데, LPC 를 이용하여 추정된 스펙트럼 포락선은 실제 음성신호를 잘 표현해 주지는 못한다. 그래서 스펙트럼의 피크들을 잘 따라가는 것으로 알려져 있는 Improved Cepstrum 방법을 이용하여 얻어진 스펙트럼 포락선을

음정변환에 이용하였다. 시간축 변환시에는 실제 음성의 발생방법을 고려하여 spectral distance 를 매 프레임마다 구해서, 그 값이 작은 프레임이 정적인 구간이고 그 값이 큰 부분은 실제 음성이 변화하고 있는 구간이라고 판단하여, 서로 다른 시간축변환 비를 적용하였다.

음성 세그멘트들을 연결할 때 생기는 피치, 위상, 그리고 스펙트럼 포락선의 불일치를 극복하기 위해서, 피치 변경, 피치 펄스 정렬(pitch pulse alignmemt) 그리고 spectral smoothing 과정을 수행하였다. 피치 펄스 정렬 시에는 기본주파수 성분이 서로 동 위상으로 더해질 수 있도록 시간축을 이동함으로써, 위상차이를 극복할 수 있었다. 스펙트럼 포락선의 불일치는 양쪽 스펙트럼을 적절히 가중치를 두어 더해줌으로써 보상해 주었다.

II. ABS/OLA Sinusoidal 모델

ABS/OLA 방법에서는 음성신호가 짧은 구간에 대해서는 정적이라는 가정하에 음성을 프레임 단위로 overlap-add 되도록 나누고 각 프레임을 일정한 정현파들의 합으로 모델링한다. 추정된 신호 $\tilde{s}[n]$ 은 다음과 같이 표현된다.

$$\tilde{s}[n]=\sigma[n]\sum_{k=-\infty}^{\infty}w_s[n-kN_s]\tilde{s}^k[n-kN_s] \tag{1}$$

여기서 $\sigma[n]$ 은 신호의 에너지 포락선이고 N_s 는 합성구간의 길이 이다. 그리고 $w_s[n]$ 는 overlap-add 됐을 때 전 구간에 대해서 그 크기가 1 이 되는 특성을 갖는 윈도우 함수를 사용하였다.

$$\sum_{k=-\infty}^{\infty}w_s[n-kN_s]=1 \tag{2}$$

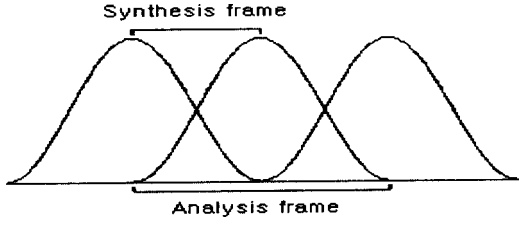

그림 1. Hanning 윈도우를 이용한 OverLap-Add Model

그리고 매 프레임마다의 합성신호 $\tilde{s}^k[n]$ 는 적절한 수의 정현파들의 합으로 나타내어진다.

$$\tilde{s}^k[n]=\sum_{j=1}^{J[k]}A_j^k\cos(\omega_j^k n+\phi_j^k) \tag{3}$$

여기서 A_j^k , ω_j^k , 그리고 ϕ_j^k 는 각각 정현파의 크기, 주파수, 그리고 위상을 나타낸다. 이 각각의 파라메터들은 Analysis-by-Synthesis 방법을 통하여 반복적으로 구하게 된다[2][3].

III. ABS/OLA system 을 이용한 음성변환

3.1 음성변환을 위한 합성 모델

음성신호의 변환을 수행하기 위해서는 피치 정보를 담고 있는 형태로 적절히 변형된 모델이 필요하다. 이러한 목적에 의해 음성신호를 다음과 같이 quasi-harmonic 형태로 모델링 할 수 있다.

$$\begin{aligned}\tilde{s}^k[n]&=\sum_{j=1}^{J[k]}A_j^k\cos[(j\omega^k{}_o+\Delta_j^k)n+\phi_j^k]\\&=\sum_{j=1}^{J[k]}A_j^k[\cos(\Delta_j^k n)\cos(j\omega_o^k n+\phi_j^k)-\sin(\Delta_j^k n)\sin(j\omega_o^k n+\phi_j^k)]\end{aligned} \tag{4}$$

여기서, $J[k]$는 $J[k]\omega_0^k\le\pi$를 만족하는 가장 큰 정수이고, Δ_j^k는 기본 주파수의 배음 성분과 ABS 과정에서 실제로 찾은 정현파의 주파수 값의 차이를 나타내는 부분이다. 이 값은 일반적으로 매우 작으므로 각각의 배음성분에 독립적으로 작용하는 진폭변조로 볼 수 있다. 시간축변환비를 ρ_k , 음정변환비를 β_k 라 할 때, 한 프레임의 음성 신호에 대한 합성식은 다음과 같다.

$$s[n+N_s]=\sigma\left[\frac{n}{\rho^k}+kN_s\right]\left\{w\left[\frac{n}{\rho^k}\right]\tilde{s}^k_{\rho^k,\beta^k}[n]+w\left[\frac{n}{\rho^k}-N_s\right]\tilde{s}^{k+1}_{\rho^{k+1},\beta^{k+1}}[n-\rho^k N_s]\right\} \tag{5}$$

이 때

$$\tilde{s}^k_{\rho^k,\beta^k}[n]=\sum_{j=1}^{J[k]}A_j^k\cos[j\beta_k\omega_o(n+\delta^k)+\frac{\Delta_j^k n}{\rho^k}+\phi_j^k]$$

$$\tilde{s}^{k+1}_{\rho^{k+1},\beta^{k+1}}[n]=\sum_{j=1}^{J[k+1]}A_j^{k+1}\cos[j\beta_{k+1}\omega_o(n+\delta^{k+1})+\frac{\Delta_j^{k+1}n}{\rho^k}+\phi_j^{k+1}]$$

이고, δ^k 는 global time shift 로써 시간축 혹은 음정 변환을 수행한 뒤에 프레임간에 위상이 맞지 않는 부분을

보정해 주는 값이다 [2].

3.2 Improved Cepstrum[4]을 이용한 음정변환

음정 변환시에는 두 가지 고려해야 할 사항이 있는데 그 중 하나는 음색에 관한 문제로서, 음성신호를 ABS/OLA 방법을 통해 분석하면 스펙트럼 영역에서 중요한 성분들을 찾아내어 그 성분들의 진폭, 주파수, 위상을 가지고 있는 것이기 때문에 스펙트럼 포락선 정보까지 가지고 있게 된다. 이 데이터를 가지고 그대로 음정변환을 수행하면 음성신호의 스펙트럼 포락선도 같이 변하므로 음색이 변하게 된다. 다음 그림은 스펙트럼 포락선을 보존하지 않고 갖고 있는 배음성분들을 그대로 움직여 음정변환한 경우를 보여준다. 이 때는 음색이 함께 변하게 된다.

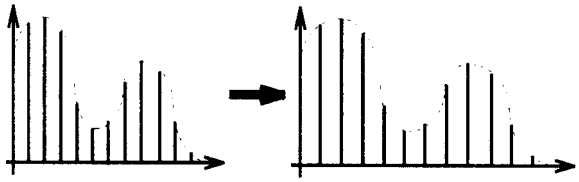

그림 2. 스펙트럼 포락선과 주파수가 같이 변한 경우

또 다른 하나의 문제점은 음정변환비가 1 보다 작은 경우에 합성음은 저주파 쪽으로 압축됨으로 고주파 성분이 없어져 음질저하를 초래하게 된다는 것이다. 그러므로 음색을 바꾸지 않고 음정만을 바꾸고자 할 때는 스펙트럼 포락선을 유지하면서 기본주파수만을 바꿀 수 있어야 한다. 그러기 위해서는 먼저 음성신호의 스펙트럼 포락선을 추정하고, 이 스펙트럼 포락선을 이용하여 배음성분들이 가지고 있는 포락선 정보를 제거한다. 그리고 남은 신호를 가지고 음정변환을 수행한다. 이때 음정변환비가 1 보다 작아서 고주파성분이 없어질 경우에 대비해 기존의 배음성분들을 interpolation 한 다음 변환된 주파수에서 resampling 하는 방식으로 음정변환을 수행한다. 그러고 난 다음 다시 포락선 정보를 얹어 줌으로 해서 음색이 보존된 음정변환을 수행할 수 있다. 그림 3 은 스펙트럼 포락선을 보존하면서 배음성분들의 위치만 변환하는 개략도 이다.

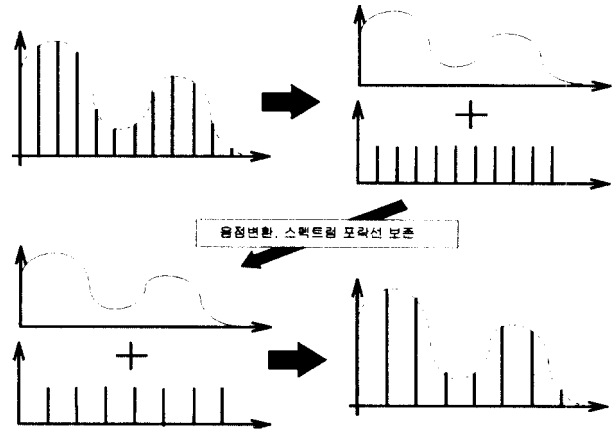

그림 3. 스펙트럼 포락선을 유지시키는 음정변환 과정

본 논문에서는 보다 정확한 스펙트럼 포락선을 추정하기 위하여 improved cepstrum 방법을 이용하였다[4]. 이 방법은 LPC 에 의해 추정된 스펙트럼 포락선에 비해 원 음성 신호의 스펙트럼 피크들을 훨씬 더 잘 이어주므로, 음정변환된 후에도 원 음성의 스펙트럼 포락선을 잘 유지시켜 보다 원음에 가깝고 선명한 음성을 합성해 낼 수가 있다.

3.3 가변적 시간축 변환

ABS/OLA Sinusoidal model 에서는 매 프레임마다 음성신호를 표현하는 정현파들의 성분들을 추출하여 그 값들을 가지고 있으므로, 그 성분들을 이용하여 우리가 원하는 길이만큼을 매 프레임마다 합성해낼 수 있고 합성된 그 신호를 overlap-add 함으로써 시간축 변환된 음성을 얻을 수 있다. 그러나 모든 프레임에 대해서 동일한 크기로 시간축을 변화시켰을때는 무성음 구간이나 음성이 변화하고 있는 구간에 대해 음질저하 현상이 나타나게 된다. 그리고 실제의 음성신호도 모든 구간에 대해 동일한 크기로 시간축이 변환되지는 않는다[6]. 음성이 변화되고 있는 구간에 대해서는 그 길이가 비슷하게 유지되고, 안정되고 소리가 변하지 않는 구간에 대

해서 그 길이가 상대적으로 크게 변하게 된다. 그러므로 실제로 시간축을 변환할 시에는 이러한 점들을 고려하여 부분적으로 다른 시간축 변환비를 사용하여야 한다. 본 논문에서는 현재 프레임을 기준으로 좌우 프레임 사이의 스펙트럼 포락선 차이를 계산하여 이 값이 작은 프레임이 정적인 구간이라 가정하고, 이 부분들을 보다 우선적으로 시간축 변환을 하도록 하였다. 그리고 이런 구간에 대해서는 2 배 정도까지 시간축을 늘이거나 0.5 배 정도로 줄였을 때, 음질 저하가 크게 나타나지 않았다. 그래서 시간축을 늘이는 경우에는 스펙트럼 포락선의 차이가 작은 순으로 시간축이 2 배까지 확장되도록 하여 원하는 길이까지 시간축 변환을 수행하도록 하였고, 길이를 줄이는 경우는 0.5 배로 하여 수행하였다. 이러한 가변적 시간축변환을 적용하여 동일한 크기로 시간축을 변환했을 때에 비해 음질저하 현상을 많이 줄이고 보다 좋은 소리를 합성해 낼 수 있었다.

IV. 연결 음성합성

연결 합성할 두 음성 단위들은 서로 다르게 발화된 음성으로부터 얻어진 것들이므로, 이 둘 사이에는 시간영역 혹은 주파수 영역에서의 신호 특성에 커다란 차이가 존재한다. 크게 피치, 위상, 스펙트럼의 차이를 들 수 있는데 피치 차이는 연결할 두 세그멘트의 피치를 같게 변환함에 의해서 쉽게 극복할 수 있다[5].

4.1 피치 펄스 정렬(Pitch pulse alignment)

음성신호를 피치 펄스 열이라고 본다면 프레임과 프레임 사이의 피치 펄스의 위치를 적절히 배열하는 것은 위상 차이를 보상해 주는데 필수적이고, 이 피치 펄스 정렬(pitch pulse alignment)은 연속적인 프레임들을 overlap-add 하기 전에 sinusoid component 의 phase 를 조절함에 의해서 이루어 질 수 있다. 다음 그림은 onvelap-add 하기 전의 두 프레임의 pitch pulse 들의 관계를 도식적으로 보여주고 있다. 이 pitch pulse 들의 위치는 pitch pulse onset time τ_k 와 τ_{k+1}, 음정 변환비 β_k 와 β_{k+1}, 피치 주기 T_0^k 와 T_0^{k+1} 그리고 이전 프레임에서 계산되었던 time shift δ^k 에 의해 결정된다.

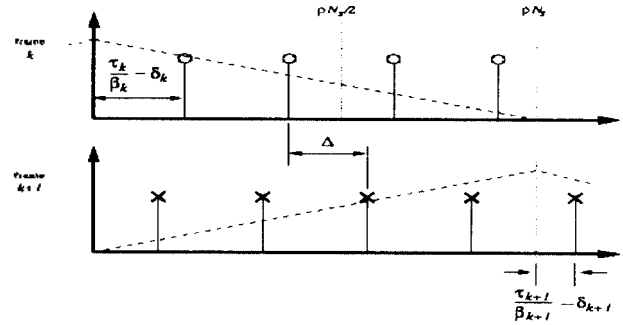

그림 4. 시간축과 피치를 변환한 후의 피치펄스모양[5]

Pitch pulse alignment 는 두 프레임의 피치 펄스의 위치를 맞추어 신호의 모양이 동일한 위상으로 overlap-add 될 수 있도록, 즉 overlap 되는 프레임의 중심에 인접한 두 피치 펄스사이의 간격 Δ가 최소가 되도록 k+1 번째 프레임의 pitch pulse 를 적절히 이동시킴에 의해서 이루어 질 수 있다. 그리고 피치 주기만큼의 이동은 같은 위치를 의미하므로 time shift 에 관한 식은 다음과 같이 유도될 수 있다.

$$\delta^{k+1} = \delta^k + \frac{\tau_{k+1}}{\beta_{k+1}} - \frac{\tau_k}{\beta_k} + \rho N_s \qquad (6)$$

본 논문에서는 pitch pulse onset time τ_k 와 τ_{k+1} 를 기본주파수에 해당하는 정현파가 최대값으로 되는 위치로 두었다. 즉, 기본주파수에 해당되는 파형이 서로 동 위상으로 더해질 수 있도록 함에 의해서 위상차이를 극복하였다.

4.2 Spectral Smoothing

음성신호의 연결 경계에서 생기는 또 다른 차이는 sinusoidal component 의 크기에 의해 나타내어지는 신호의 스펙트럼 모양의 불일치이다. 연결부분에서의 스펙

트럼 차이는 연결 경계근처에서의 몇 프레임에 대해 스펙트럼 포락선이 서서히 다음 세그멘트의 스펙트럼 포락선으로 변해가도록, 현재의 스펙트럼과 연결할 상대편의 스펙트럼 포락선 사이에 적절히 가중치를 두어 더한 값을 이용함으로써 보상해 줄 수 있다. 이 때, 유성음과 무성음이 합해지는 경우 등의 일부 상황하에서는 이러한 보상이 오히려 음질 저하를 초래하기도 하므로, 적절한 구간 선정이 필요하다.

4.3 연결 음성합성 실험 및 결과

국어공학센터의 2000 어절 DB[7]에서, '초성+중성'/'종성'으로 합성단위를 구성하여 음성 합성 실험을 수행하였다. 다음은 그 한 예로 '우리 실험실은 좋은 실험실이다' 라는 문장의 파형과 스펙트로그램 이다.

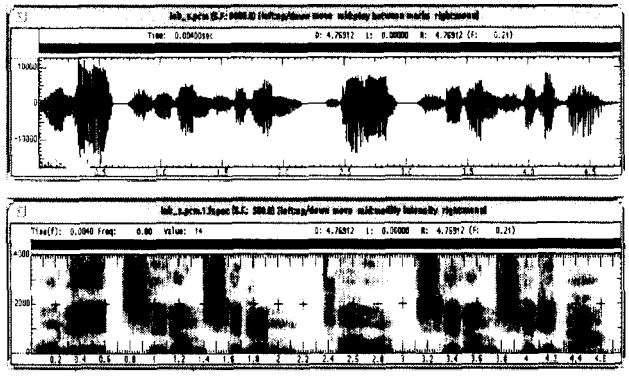

그림 5. 연결음성 합성의 한 예

이상의 방법들을 통해 앞에서 말한 연결합성시에 생기는 문제점들을 잘 극복하여 자연스럽고 매끄럽게 연결되는 합성음을 만들어 낼 수 있었다. 그러나 아직도 다른 크기의 에너지를 가지는 세그멘트들을 이어 붙임에 의해 생기는 에너지 차이등에 의한 음질저하 현상에 대한 보상 등이 필요하다.

V. 결론

본 논문에서는 ABS/OLA Sinusoidal Model 을 이용하여 시간축 및 음정 변환을 수행함에 있어서, 음성신호의 특징을 이용한 가변적인 시간축 변환과 음성신호의 스펙트럼 포락선을 잘 추정한다고 알려져 있는 Improved Cepstrum 을 음정변환에 적용하여 보다 고음질의 합성음을 만들어 낼 수 있었다. 그리고 양쪽 프레임의 기본주파수 성분이 동일한 위상으로 더해지도록 시간축을 이동하여 합성함으로써 인접 음과의 경계부분에서 발생하는 위상 차이를 해결하였다. 본 논문에서 제안한 이러한 방식을 통해 기존의 방법에 비해 우수한 성능의 합성결과를 얻을 수 있었다.

참고문헌

[1] R. J. McAuley and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," IEEE Trans. on ASSP, vol. 34, pp.744-753, Aug., 1986.

[2] E. B. George and M. J. T. Smith, "Analysis-by-Synthesis/Overlap-Add Sinusoidal Modeling Applied to the Analysis and Synthesis of Musical Tones," J. Audio Eng. Soc., Vol. 40, No. 6, pp. 497-516, June, 1992.

[3] E. B. George and M. J. T. Smith, "Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model," IEEE Trans. on SAP, Vol. 5, No. 5, pp. 389-406, Sep., 1997.

[4] S. Imai and Y. Abe, "Spectral Envelope Extraction by Improved Cepstral Method," IECE Trans. J62-A, pp. 217-223, 1979. (in Japanese)

[5] Michael W. Macon and Mark A. Clements, "Speech Concatenation and Synthesis Using an Overlap-Add Sinusoidal Model," ICASSP, Vol. 1, pp 361-364, 1996.

[6] Sung Ju Lee and Hyung Soon Kim, "Variable Time-Scale Modification of Speech Using Transient Information," ICASSP, pp 1319-1322, 1997.

[7] 김봉완 외 5명, "공동이용을 위한 단어음성DB의 구축 및 PBS 설계에 관한 검토," 제13회 음성통신 및 신호처리 워크샵 논문집, pp 256-261, 1996.