1999년도 한국음향학회 학술발표대회 논문집 제18권 제1(s)호

ABS/OLA Sinusoidal 모델에서 위상계승을 이용한 단위음성의 연결

배재현°, 변효진, 오영환
한국과학기술원 전산학과

Speech Unit Concatenation by Phase Succession in an ABS/OLA Sinusoidal Model

Jae-Hyun Bae°, Heo-Jin Byeon, Yung-Hwan Oh
Dept. of Computer Science,
Korea Advanced Institute of Science and Technology
E-mail: jhbae@bulsai.kaist.ac.kr

요약

본 논문에서는 중첩가산 Sinusoidal 합성방식에서 매칭된 정현파별로 위상을 계승하는 단위음성 연결방법을 제안한다. 선행 단위음의 마지막 프레임, 후행 단위음의 첫 프레임, 후행 단위음의 나머지 프레임의 단계로 나누어 각 단계마다 제안한 방식으로 선행 프레임의 위상을 계승하였다. 실험결과 후행 단위음의 연결 위치를 이동하는 기존의 방식을 사용한 연결음에 비해 연결부분에서 음성파형의 급격한 변화가 줄었다.

1. 서론

Sinusoidal 모델은 음성신호를 주파수 영역에서 공명이 일어나는 각 정현파의 주파수와 크기, 위상 등으로 표현하는 모델이다. Sinusoidal 모델에 기반한 합성방식은 음성을 분석하여 Sinusoidal 모델에 필요한 파라메타를 추출하고, 합성시에는 추출한 파라메타를 이용하여 음성을 재구성하게 된다. 이 방식은 음성을 파라메타화하여 갖고 있으므로 재합성시 음성 특성의 조절이 용이하며, 합성음의 음질 또한 우수한 장점을 가진다[1]. Sinusoidal 모델을 이용한 합성방식에는 중첩가산법을 이용한 합성방식과 [2] 다차원 다항식을 이용하여 인접한 프레임들을 서로 보간하는 방법 등이 있다[1]. 다차원 다항식을 이용한 보간법은 프레임 사이를 1차원, 3차원 다항식을 이용하여 보간하며 프레임마다 다항식의 계수를 다시 구해야 하기 때문에 계산량이 많다. 이에 반해 중첩가산 방식을 이용한 합성방식은 합성된 프레임들에 가중치를 두어 중첩가산하는 방식으로 다차원 다항식을 이용하여 보간, 합성하는 방법에 비해 계산량이 적다는 장점이 있다[2], [3].

중첩가산 Sinusoidal 모델을 이용한 TTS 시스템(Text-To-Speech System)은 합성음을 구성하는 각각의 단위음성에 대해 필요한 파라메타를 추출하여 저장하고, 음성 합성시에는 저장된 파라메타를 이용하여 재합성한 각각의 단위음을 연결하여 합성음을 구성한다. 그러나 단위음성에서 추출한 정보들은 같은 음소라도 녹음 환경에 따라 차이를 보인다. 특히 위상정보는 많은 차이를 보이게 되므로 단위음 연결시에는 단위음간의 서로 다른 위상을 보간하여야 한다. 보간은 피치펄스 정렬(pitch pulse alignment)을 이용해 후행 단위음을 선행 단위음의 기본 주기 간격에 맞추어 연결함으로써 단위음간의 위상왜곡을 줄이는 방식이 많이 사용된다[4]. 그러나 후행 단위음성을 이동하여 연결하는 방법은 각 정현파 위상을 조정하는 것이 아니라 전체적인 기본 주기만 고려하여 음성파형 전체를 이동하고 각각의 정현파에 대해서는 보간을 하지 않는다. 따라서 연결음의 파형은 기본 주기 간격만이 유지될 뿐이고, 연결부분에서의 파형이 서로 다르다는 문제점이 생기게 된다. 본 논문에서는 각각의 정현파에 대해 위상정보를 조정함으로써 이러한 왜곡을 줄이는 방법을 제시하고 실험결과를 보인다.

본 논문의 구성은 다음과 같다. 2장에서 Sinusoidal 모델과 중첩가산 합성방식에서의 단위음성 연결에 대해 설명하고, 3장에서는 기본적인 중첩가산방식에서 사용하는 단위음성 연결방식의 문제점을 최소화할 수 있는 선행 단위음의 위상을 계승방법을 제안한다. 4장에서는 제안한 방법에 대한 실험과 결과를 설명하고 5장에서 결론을 맺는다.

2. Sinusoidal 모델에서의 중첩가산 합성 방식과 단위음성의 연결

Sinusoidal 모델은 식 (1)과 같이 음성신호가 서로 다른 주파수, 크기, 위상을 갖는 정현파의 합으로 이루어진다는 가정 하에 음성신호를 표현한다[1].

$$s^k[n] = \sum_{j=1}^{J[k]} A_j^k \cos(n\omega_j^k + \Phi_j^k) \qquad (1)$$

식 (1)에서 A_j^k, ω_j^k, Φ_j^k, $J[k]$는 각각 k번째 프레임에서 j번째 정현파의 크기, 주파수, 위상 그리고 정현파의 개수를 나타낸다. Sinusoidal 방식에 기반한 중첩가산 합성법은 원하는 길이와 주파수의 합성음을 얻기 위해 추출된 단위음의 파라메타를 β, ρ를 사용하여 식 (2)에서와 같이 변환하여 합성한다.

$$\hat{s}^{\,k}_{\rho_k,\beta_k}[n] = \sum_{j=0}^{J[k]} A_j^k \cos[j\beta_k\omega_0^k(n+\delta^k) + \frac{\Delta_j^k}{\rho_k} + \Phi_j^k] \qquad (2)$$

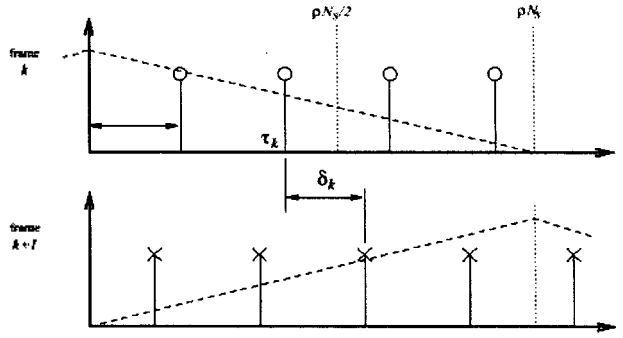
식 (2)에서 Δ 는 기본주파수의 정수배와 실제 정현파의 주파수의 차이를 나타내고, β, ρ는 각각 주파수 영역에서의 기본주파수 변환비, 시간 영역에서의 지속시간 변환비를 나타낸다. δ는 β, ρ 등에 따른 변환으로 생긴 위상의 불일치를 해소하기 위해 합성음이 이동하여야 할 시간축 상의 샘플 수를 나타내며, k는 프레임의 인덱스이다.

식 (2)를 이용하여 합성된 프레임들은 식 (3)과 같이 합성된 두 프레임 각각에 창함수를 이용해 적절한 가중치를 곱하여 중첩가산한다[2].

$$\begin{aligned} s[n+N_s] &= w[n/\rho^k]\,\hat{s}^{\,k}_{\rho^k,\beta^k}[n] \\ &\quad + w[n/\rho^k - N_s]\,\hat{s}^{\,k+1}_{\rho^{k+1},\beta^{k+1}}[n-\rho_k N_s] \end{aligned} \qquad (3)$$

식 (3)에서 w는 중첩에 사용하기 위해 합성된 프레임에 곱하는 창함수이고, N_s는 합성구간이다.

추출된 파라메터와 식 (2), (3)을 이용해 합성된 단위음들은 독립적으로 녹취된 것들이므로 위상차이가 생기게 된다. 따라서 단위음의 중첩가산연결시, 단위음간의 보간과정이 필요하게 된다. 기존의 Sinusoidal 중첩가산방식의 경우 pitch onset time에서 모든 정현파의 위상이 같아진다는 가정 하에 pitch onset time을 이용하여 후행 단위음을 시간축 상에서 이동시켜 연결하는 방법이다. 이 방식은 단위음성간의 연결시 [그림 1]에서 보는 바와 같이 연결되는 프레임간에는 δ 만큼 피치펄스의 간격차가 생기게 된다.따라서 식 (4)와 같이 시간축 상에서 후속 프레임을 δ만큼 이동시켜 연결하거나 그에 해당하는 위상만큼 후행 프레임의 위상을 변화시켜 연결한다. 나머지 프레임에 대해서는 선행 프레임의 δ를 이용하여 이동할 거리를 결정한다[4].



[그림 1] 중첩가산시의 피치펄스 간격의 차.

$$\delta^{k+1} = \delta^k + \frac{\tau_{k+1}}{\beta_{k+1}} - \frac{\tau_k}{\beta_k} + \rho N_s \qquad (4)$$

식 (4)에서 τ 는 pitch onset time을 나타낸다. 그리고 식 (4)에서 구해진 δ를 식 (3)에 적용시켜 단위음성을 연결한다.

3. 위상계승을 통한 단위음성의 연결

피치펄스 정렬을 이용한 단위음 연결 방식에서는 음성파형 전체를 이동시키므로 연결부분에서 전체적인 파형은 잘 연결되지만 파형의 모양이 급격히 바뀌게 된다. 그리고 실제음성을 구성하는 정현파의 주파수가 기본주파수의 정수배가 되지는 않는다. 이러한 사실을 바탕으로 한 quasi harmonic 모델에서는 harmonic 모델에 기반한 pitch onset time에서 모든 정현파의 위상이 동일하다는 가정을 그대로 적용시키기에는 무리가 있다. 또한 pitch onset time을 계산할 때 그 과정이 부정확할 수 있다는 문제점이 있다[5]. 본 논문에서는 피치펄스 정렬을 통한 단위음성 연결방식의 문제점을 줄이기 위해 정현파간의 매칭을 통해 매칭되는 정현파 각각에 대해 위상을 계승하여 연결하는 방법을 제안한다.

후행 단위음의 첫 프레임에서는 선행 단위음의 위상을 계승하고, 이후의 프레임에서는 기존의 정현파 매칭 알고리즘을[1] 이용하여 선행 프레임과의 정현파 매칭과정으로 얻어진 정현파들의 위상에 앞 프레임에서의 위상차를 더해준다. 제안한 방법은 세 단계로 나뉘어진다. k를 선

행 단위음의 정보를 사용한 마지막 합성프레임의 인덱스라고 하면 후행 단위음의 정보가 처음 사용된 합성프레임의 인덱스는 $k+1$이 된다. k번째 합성 프레임에서는 사용된 정현파의 주파수, 위상을 식 (5), (6)과 같이 보관하여, $k+1$번째 합성 프레임에서 후행 단위음과의 연결시 정현파 매칭과 위상계승에 사용한다.

$$\hat{\Phi}_i^{\,k} = i\beta_k\omega_0^k(N_s/2+\delta_k) + \frac{\phi_i^k \Delta^k}{\rho} N_s/2 \qquad (5)$$

$$\hat{\omega}_j^{\,k} = \beta_k\omega_j^k \qquad (6)$$

식 (5), (6)에서는 단위음 합성시 사용하는 정현파의 위상과 주파수를 나타내고, $\hat{\Phi}_j^{\,k}$, $\hat{\omega}_j^{\,k}$는 단위음 연결을 위해 조정된 단위음의 위상과 주파수 정보이다. $k+1$번째 합성 프레임에서는 후행 단위음의 첫 프레임의 기본주파수를 k번째 합성 프레임의 기본주파수와 일치하도록 조정한 후 저장한 $\hat{\Phi}_j^{\,k}$, $\hat{\omega}_j^{\,k}$과 매칭한다. 그리고 $k+1$번째 합성 프레임에서 각 정현파의 위상은 매칭 결과에 따라 식 (7), (8)과 같이 결정한다.

$$\hat{\Phi}_j^{\,k+1} = \begin{cases} \hat{\Phi}_i^{\,k} & , \text{ if } (i,j) \text{ is matched} \\ \Delta_\phi + \Phi_j^{k+1} & , \text{ otherwise} \end{cases} \qquad (7)$$

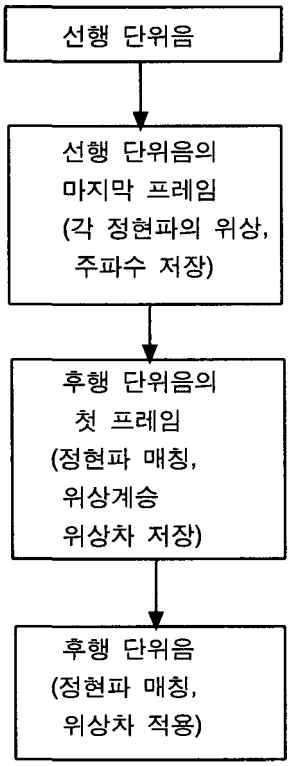
$$\Delta_\phi = \frac{1}{N_{match}} \sum_{i=1}^{N_{match}} (\hat{\Phi}_i^{\,k} - \Phi_j^{k+1}) \qquad (8)$$

식 (7)과 같이 매칭되는 정현파들은 식 (5)에 의해 구한 위상을 계승하게 되며 매칭되지 않은 정현파들에 대해서는 매칭되는 정현파들의 평균 위상변화량을 더해준다. 따라서 한 프레임 내의 정현파 간의 위상 관계도 일정하게 유지하게 된다.

$k+2$번째 합성 프레임부터는 식 (7)에서 매칭된 정현파들에 대해서만 그 정현파들이 birth-and-death 법칙에[1] 의해 소멸될 때까지 매칭함수를 적용한다. 또한 매칭되는 정현파들에 대해서 이전 프레임의 위상을 직접 계승하는 것이 아니라, 식 (9)와 같이 $k+1$번째 합성 프레임에서의 정현파 위상 변화량을 반영시킨다.

$$\hat{\Phi}_j^{\,k+m} = \begin{cases} \Phi_{1,i} - \Phi_i^{k+1} + \Phi_j^{k+m} \\ \quad , \text{ if } (i_k, j_{k+m}) \text{ is matched} \quad , \text{where } m \geq 2 \\ \Delta_\phi + \Phi_j^{k+m} \\ \quad , \text{ otherwise} \end{cases} \qquad (9)$$

식 (9)에 의하면 정현파 쌍 (i_k, j_{k+m})은 정현파 i_k가 $k+1$번째 합성 프레임 이후의 프레임들을 거치면서 소멸되지 않고 연결되어 후행 $k+m$번째 프레임의 정현파 j_{k+m}과 매칭될 때 유효하다. (i_k, j_{k+m})가 매칭되면 위상차 $\Phi_{1,i} - \Phi_i^{k+1}$를 적용하고, (i_k, j_{k+m})가 매칭되지 않으면 $k+1$번째 프레임에서의 평균 위상차 Δ_ϕ를 적용한다. 따라서 단위음 연결부분에서 매칭되는 정현파는 birth-and-death 법칙에 따라 소멸될 때까지 식 (5)-(9)식에 의해서 연결 합성음에서 그 영향을 미치게 된다. 따라서 피치펄스 정렬을 통한 연결보다 더 완만하게 과형의 모양이 후행 단위음의 모양으로 변하게 된다.

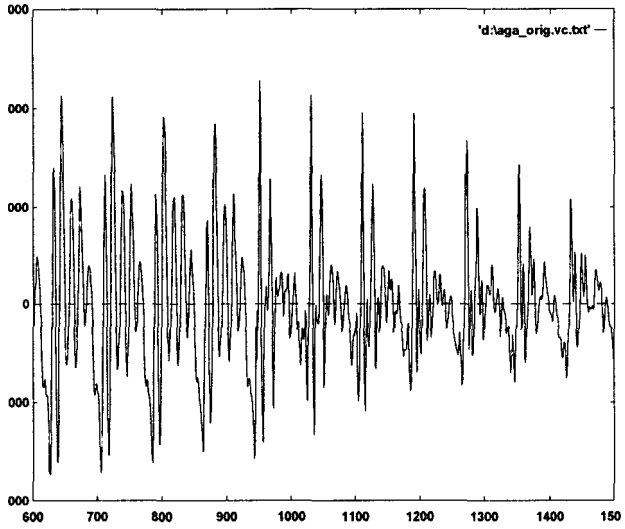


[그림 2] 위상 계승법을 이용한 단위음 연결의 블록도.

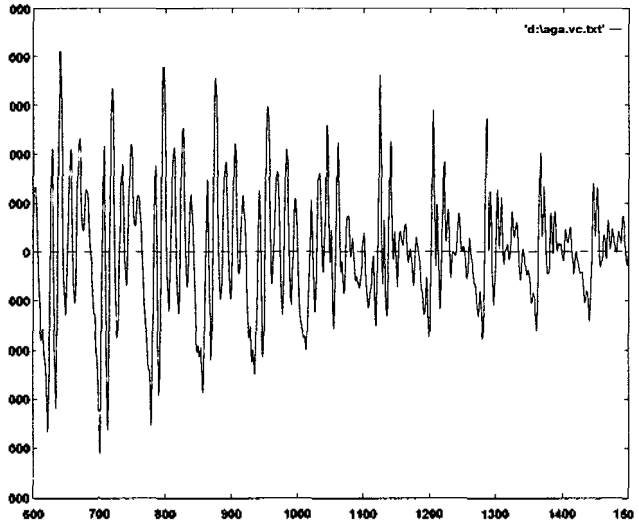
4. 실험 및 결과

실험에 사용한 단위음성들은 기존 TTS시스템[8]에 사용되는 VCV(Vowel-Consonant-Vowel) 연쇄음을 기반으로 한 단위음성 데이터베이스에서 추출한 것이다. 이 단위음성 데이터베이스는 16kHz로 샘플링 되어있으며, 전문 여성 아나운서가 발성한 음성에서 추출되었다. 비교실험을 위해 기존의 피치펄스 정렬방식을 이용한 단위음 연결방식과 본 논문에서 제안한 방식을 구현하였다. 파라메타 추출에는 2048 포인트 FFT (fast fourier transform)와 SEEVOC (spectral envelope estimation vocoder) 알고리즘을[6] 사용하였다. 분석 프레임은 20 ms단위로 하였고,

10 ms씩 프레임을 이동하면서 분석하였다. 합성은 20 ms 단위 프레임들을 10 ms씩 중첩가산하였다. [그림3]에서 보는 바와 같이 파형의 연결시 기존의 방법에서는 단위음간의 기본주파수의 연결은 자연스러우나 파형의 모양이 급격히 변함을 알 수 있다. 제안한 방법이 적용된 연결음은 기본주파수의 연결은 물론 파형이 점차적으로 변해감을 알 수 있다. 또한 pitch onset time을 사용하지 않으므로 계산량이 줄어들고, pitch onset time 계산의 부정확함에서 오는 문제점도 피할 수 있게 된다.



(a) 기존의 방법을 이용한 중첩가산



(b) 위상계승을 통한 중첩가산

[그림3] ' ㅏ'와 ' ㅏ가'의 연결음의 파형.

5. 결론

본 논문에서는 ABS/OLA Sinusoidal 모델에서 사용되는 피치필스 정렬을 통한 단위음성 연결방식에서 합성음의 파형이 연결부분에서 급격히 변하는 것과 pitch onset time을 계산하여야 하는 문제점을 해결하기 위하여 단위음성간의 정현파 매칭을 통하여 매칭된 정현파의 위상을 후행 프레임에서 계승하는 방법을 제안하였다. 본 논문에서 제안한 방법으로 단위음을 연결한 결과 자연스럽게 연결된 합성음을 만들 수 있었으며, 이러한 방법을 음성합성시스템에 적용할 수 있을 것이다.

참고문헌

[1] R. J. McAulay 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 Speech and Audio Processing, vol. 5, no. 5, pp. 389-406, september, 1997

[4] M. W. Macon and M. A. Clements "Speech Concatenation and Synthesis Using An Overlap-Add Sinusoidal Model," ICASSP, vol. 1, pp. 361-364, 1996

[5] R. J. McAulay and T. F. Quatieri, "Pitch Estimation and Vocing Detection Based on a Sinusoidal Speech Model," ICASSP, pp. 249-252 Apr. 1990

[6] "The spectral envelope estimation vocoder," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-29, no. 4, pp. 786-794, Aug. 1981

[7] 구자형, 최무열, 김형순 "Analysis-By-Synthesis / Overlap-Add(ABS/OLA) sinusoidal Model을 이용한 음성변환과 연결음성 합성", KSCSP '98 15권 1호 pp.. 339-343, 1998

[8]"W/S용 Text-to-Speech 시스템 기술개발에 관한 연구". 한국과학기술원, 산업자원부, 11. 1998