1999년도 한국음향학회 학술발표대회 논문집 제18권 제2(s)호

ABS/OLA Sinusoidal 모델을 이용한 문서-음성 변환시스템의 구현

배 재 현○ 변 효 진 오 영 환

한국과학기술원 전산학과

Implementation of Text-to-Speech System using ABS/OLA Sinusoidal Model

Jae-Hyun Bae○ Heo-Jin Byeon Yung-Hwan Oh

Department of Computer Science

Korea Advanced Institute of Science and Technology

jhbae@bulsai.kaist.ac.kr

요 약

본 논문에서는 중첩가산 Sinusoidal 합성방식에서 위상계승에 의한 단위음 연결법과 다프레임간 정현파 크기의 보간법을 제안한다. 그리고 합성 프레임의 중심이 pitch onset time이라고 가정하고, 음성에서 분리한 성도 모델의 위상을 음성 전체의 위상으로 사용하는 방법을 제안한다. 제안한 방법으로 문서-음성 변환 시스템(Text-to-Speech System, TTS System)을 구현한 결과 단위음 연결시 연결부분의 파형 왜곡이 감소함을 알 수 있었고, 부드럽게 연결된 합성음을 얻을 수 있었다.

1 서론

Sinusoidal 모델은 음성신호를 주파수 영역에서 공명이 일어나는 각 정현파의 주파수와 크기, 위상 등의 파라메타로 표현하는 모델이다. 중첩가산 Sinusoidal 모델을 이용한 문서-음성 변환 시스템은 합성음을 구성하는 각각의 단위음에 대해 필요한 파라메타를 추출하여 저장하고, 음성 합성시에는 저장된 파라메타를 조절하여 재합성한 단위음을 연결하여, 합성음을 구성한다.

기존의 합성방법에서는 음성 파라메타를 음원과 성도 모델로 분리하여 변환한 후, 합성음을 생성한다[2]. 그러나 음원과 성도모델의 정현파 변환에 사용하는 pitch onset time은 정확히 측정하기가 어려우므로, 측정의 오차에 따른 음성의 왜곡이 발생한다[7, 8]. 그리고 합성한 단위음의 연결시에는 두 단위음의 위상과 크기를 보간하여야 하는데, 기존의 단위음 연결법에서는 두 단위음의 위상을 보간하기 위해, 피치펄스 정렬(pitch pulse alignment)을 이용해 후행 단위음을 선행 단위음의 기본주기 간격에 맞추어 연결함으로써 단위음 간의 위상왜곡을 줄이는 방법을 사용하였다[5]. 그러나 후행 단위음을 이동하여 연결하는 방법은 합성음을 이루는 정현파 각각의 위상을 보간하는 것이 아니라, 전체적인 기본 주기만 고려하여 음성 파형 전체를 이동한다. 따라서 연결음의 파형은 기본 주기 간격만이 유지될 뿐이고, 연결부분에서 파형의 모양이 급격히 변하게 되는 문제점이 있다. 그리고 피치 펄스 정렬 방법을 사용하기 위해서 pitch onset time을 사용하는데, 단위음 생성시와 같이 측정의 오차에 따른 왜곡이 발생하는 문제점이 있다[7, 8]. 선, 후행 단위음의 크기를 보간하기 위하여 기존의 방법에서는 연결부분에 해당하는 두 프레임 사이에서 선, 후행 단위음의 크기를 보간한다. 그러나 이 경우 연결음 파형의 크기가 두 프레임 사이에서 급격히 변하게 되어 두 단위음이 자연스럽게 연결되지 못한다.

본 논문에서는 이러한 문제점들을 해결하기 위해 단위음 생성시 음성에서 성도 모델의 위상을 추출한 후, 합성 프레임 내에서 음원 모델의 위상이 동기화될 때 정현파들의 위상으로 사용하는 방법을 제안한다. 그리고 위상 계승과 다프레임간 정현파 크기의 보간을 통한 단위음 연결방법을 제안하고 합성음을 생성한다.

본 논문의 구성은 다음과 같다. 2 장에서 중첩가산 Sinusoidal 모델에서의 음성 파라메타 조절과 단위음 연결에 대해 설명한다. 그리고 3 장에서 성도 모델의 위상 분리를 통한 음성 파라메타의 조절 방법을 제안하고, 4 장에서는 선행 단위음의 위상을 계승하는 방법과, 선, 후행 단위음 연결시 다프레임간 정현파 크기의 보간법을 제안한다. 그리고 5 장에서 제안한 방법을 문서-음성 변환 시스템에 적용, 구현하고 결과를 보인 후, 마지막으로 6 장에서 결론을 맺는다.

2 ABS/OLA Sinusoidal 모델

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

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

식 (1)에서 A_j^k, ω_j^k, Φ_j^k, $J[k]$은 각각 k번째 프레임에서 j번째 정현파의 크기, 주파수, 위상 그리고 정현파의 갯수를 나타낸다. Sinusoidal 모델에 기반한 중첩가산 합성법에서는 원하는 지속시간과 주파수의 합성음을 얻기 위해, 추출된 단위음의 파라메타를 주파수 변환비(β), 지속시간 변환비(ρ)를 사용하여 식 (2)에서와 같이 변환하여 합성한다[3].

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

식 (2)에서 Δ는 기본주파수의 정수배와 실제 정현파의 주파수의 차이를 나타내고, δ는 β, ρ등에 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 각 프레임에 창함수($w[n]$)를 이용해 적절한 가중치를 곱하여 중첩가산한다[3].

$$\begin{aligned} s[n+N_s] &= w[\frac{n}{\rho^k}]\tilde{s}_{\rho_k,\beta_k}[n] \\ &\quad + w[\frac{n}{\rho^k} - N_s]]\tilde{s}^{k+1}_{\rho_k,\beta_{k+1}}[n-\rho_k N_s] \quad (3) \end{aligned}$$

식 (3)에서 N_s는 합성구간이다. 그러나 식 (2), (3)에 사용된 음성 파라메타들은 독립적으로 녹취된 단위음에서 추출한 것들이므로, 같은 음소를 표현하는 두 단위음 사이에도 위상차이가 생기게 된다. 따라서 단위음의 중첩가산연결시, 단위음간에 위상의 보간 과정이 필요하다. 기존의 Sinusoidal 중첩가산방식의 경우 pitch onset time에서 모든 정현파의 위상이 같아진다는 가정 하에, 선, 후행 단위음의 피치 간격이 일정하도록 후행 단위음을 시간축 상에서 이동시켜, 연결음의 피치 간격이 같도록 만드는 피치 펄스 정렬 방법(pitch pulse alignment)을 사용한다[5]. 따라서 시간축 상에서 후속 프레임을 피치펄스의 간격차(δ)만큼 이동시켜 연결하거나, 그에 해당하는 위상만큼 후행 프레임의 위상을 변화시켜 선행 단위음과 연결한다. 나머지 프레임에 대해서는 선행 프레임의 δ를 이용하여 이동할 거리를 결정한다[5].

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

식 (4)에서 τ는 pitch onset time을 나타낸다. 음성 합성시에는 식 (4)에서 구해진 δ를 식 (2)에 적용시켜 후행 단위음을 생성하고, 식 (3)에 의해 선행 단위음과 중첩가산하여 연결합성음을 생성한다.

3 성도 모델의 위상 분리를 통한 음성 파라메타의 조절

기존의 합성방법에서는 음성을 성도와 음원 모델로 분리해서 파라메타를 조절한다[1, 2]. 이 때 분리하는 정보는 아래 식과 같이 위상과 크기 정보이다.

$$\begin{aligned} \Phi(\omega_k) &= H(\omega_k) + \phi(\omega_k) \\ A(\omega_k) &= M(\omega_k) \cdot a(\omega_k) \end{aligned} \quad (5)$$

식 (5)에서 Φ, A는 각각 음성을 구성하는 정현파들의 위상과 크기를 나타내며, H ,M는 성도 모델의 위상과 크기를 나타낸다. 그리고 ϕ, a는 음원모델의 위상과 크기를 나타낸다. 음성을 구성하는 정현파들은 음원 모델에서 모든 정현파들은 pitch onset time에서 위상이 0이 된다는 가정 하에, 식 (6)을 이용하여 음원과 성도 모델로 분리한다[2, 7].

$$\begin{aligned} \Phi_k(n=0) &= \Psi_k(n=0) + \phi_k(n=0) \\ &\quad where,\ \ \phi_k(n=0) = \omega_k \cdot P \end{aligned} \quad (6)$$

식 (6)에서 P는 분석 프레임의 pitch onset time과 프레임 중심과의 거리이다. 피치변환된 정현파의 크기는 위에서 분리한 음원 모델과 성도 모델의 스펙트럼 궤적 상에서 재 샘플링한 정현파 크기의 곱으로 나타낸다. 그리고 변환된 정현파의 위상은 음성의 성도모델과 음원모델에서 해당하는 정현파의 위상을 더하여 나타낸다. 그러나 pitch onset time은 정확히 측정하기 어렵기 때문에, 측정의 오차에 의해 음성의 왜곡이 발생할 수 있다[7, 8].

본 논문에서는 이와 같은 문제점을 해결하기 위하여, 성도 모델의 위상을 전체 음성의 위상으로 사용하는 방법을 제안한다. 성도 모델의 위상은 식 (7)과 같이 켑스트럼을 이용하여 구할 수 있다[10].

$$\Psi_k(n=0) = -2\sum_{m=1}^{NCEP} c_m sin(m\omega_k) \quad (7)$$

식 (7)에서 $\Psi_k, M_k(n=0)$은 각각 성도모델에서 k번째 정현파의 위상과 크기를 나타내고, c_m은 켑스트럼 계수이다. 음성은 짧은 구간에서 안정한 특성을 보이므로, 한 프레임 내에서는 음성의 특징이 변하지 않는다고 가정할 수 있다[11]. 가정에 의해 음원 모델을 구성하는 정현파의 위상이 0이 되는 pitch onset time에서는, 음성 전체의 위상은 성도 모델의 정현파 위상만으로 이루어지게 된다. 그러므로 합성프레임의 중심이 pitch onset time이 되면 음원 모델의 위상을 구하지 않고, 성도 모델의 위상만으로 합성프레임의 전체 위상을 구성할 수 있다. 따라서 본 논문에서는 합성시 합성 프레임의 중심을 pitch onset time으로 가정하여 합성음을 생성한다. 음성 전체의 위상($\Phi_k(n=0)$)은 식 (6)에서 P가 0이 되어 식 (8)과 같이 성도 모델의 위상만으로 합성음의 위상을 구할 수 있다.

$$\Phi_k(n=0) = \Psi_k(n=0) \text{ ,where } P \simeq 0 \quad (8)$$

피치 변환된 정현파의 크기를 구하기 위해서 원 음성의 파워 스펙트럼에서 SEEVOC (Spectrum Envelope Estimation Vocoder)에서 사용한 방법을 이용하여 스펙트럼 포락을 구한 후[9], 각 정현파의 주파수에 해당하는 곳을 추출하여 피치변환 후의 정현파의 크기로 한다. 본 논문에서 제안한 방법으로 음성의 파라메타를 조정하면 pitch onset time측정의 오차로 인해 성도와 음원 모델로 나누는 과정에서 생기는 음성의 왜곡을 줄일 수 있다.

4 위상계승 및 다프레임간 정현파 크기 보간에 의한 단위음 연결

피치펄스 정렬을 이용한 단위음 연결 방법에서는 음성파형 전체를 이동시키므로 연결부분에서 전체적인 파형은 잘 연결되지만 파형의 모양이 급격히 바뀌게 된다. 또한 pitch onset time을 계산할 때 그 과정이 부정확할 수 있다는 문제점이 있다[7, 8]. 본 논문에서는 피치펄스 정렬을 통한 단위음 연결방

법의 문제점을 해결하기 위해, 매칭되는 정현파 각각의 위상을 계승하여 연결하는 방법과, 선, 후행 단위음 사이에서 다프레임간 정현파 크기의 보간방법을 제안한다.

4.1 위상계승을 이용한 단위음 연결

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 더한다.
제안한 방법은 세 단계로 나뉜다. k를 선행 단위음의 정보를 사용한 마지막 합성프레임의 인덱스라고 하면 후행 단위음의 정보가 처음 사용된 합성프레임의 인덱스는 $k+1$이 된다. k번째 합성 프레임에서는 사용된 정현파의 주파수($\tilde{\omega}_j^k$), 위상($\tilde{\Phi}_j^k$)을 저장하고, $k+1$번째 합성 프레임에서는 주파수에 의한 정현파 매칭을 통해 저장한 k번째 합성 프레임의 위상을 계승한다. $k+1$번째 합성 프레임에서는 후행 단위음의 첫 프레임의 기본주파수를 k번째 합성 프레임의 기본주파수와 일치하도록 조정한 후, 주파수(ω_j^{k+1})과 위상(Φ_j^{k+1})을 $\tilde{\omega}_j^k$, $\tilde{\Phi}_j^k$과 매칭한다. 그리고 $k+1$번째 합성 프레임에서 각 정현파의 위상은 매칭 결과에 따라 식 (9), (10)과 같이 결정한다.

$$\tilde{\Phi}_j^{k+1} = \begin{cases} \tilde{\Phi}_i^k + \tilde{\omega}_i^k \cdot N_{s_k} & \text{if } (i,j) \text{ is matched} \\ j\Delta_\Phi + \Phi_j^{k+1} & \text{otherwise} \end{cases} \tag{9}$$

$$\Delta_\Phi = \frac{1}{N_{match}} \sum_{j=1}^{N_{match}} \frac{\tilde{\Phi}_j^{k+1} - \Phi_j^{k+1}}{j} \tag{10}$$

식 (9)와 같이, 매칭되는 정현파들은 이전 프레임의 매칭되는 정현파의 위상을 계승하게 되며 매칭되지 않은 정현파들에 대해서는 매칭되는 정현파들의 평균 위상변화량을 더한다. 따라서 한 프레임 내의 정현파 간의 위상 관계도 일정하게 유지된다. $k+2$번째 합성 프레임부터는 식 (9)에서 매칭된 정현파들에 대해서만 그 정현파들이 birth-and-death 법칙[1]에 의해 소멸될 때까지 매칭함수를 적용한다. 또한 매칭되는 정현파들에 대해서 이전 프레임의 위상을 직접 계승하는 것이 아니라, 식 (11)과 같이 $k+1$번째 합성 프레임에서의 정현파 위상 변화량을 반영한다.

$$\tilde{\Phi}_j^{k+m} = \begin{cases} \Phi_j^{k+m} + \tilde{\Phi}_i^{k+1} - \Phi_i^{k+1} \\ \text{, if } (i_k, j_{k+m}) \text{ is matched} \\ \\ j\Delta_\Phi + \Phi_j^{k+m} \\ \text{,otherwise} \end{cases} \tag{11}$$
$$\text{, where } m \geq 2$$

식 (11)에서 정현파 쌍 (i_k, j_{k+m})은 정현파 i_k이 $k+1$번째 합성 프레임 이후의 프레임들을 거치면서 소멸되지 않고 연결되어 $k+m$번째 합성 프레임의 정현파 j_{k+m}과 매칭될 때 유효하다. (i_k, j_{k+m})이 매칭되면 위상차 $\tilde{\Phi}_i^{k+1} - \Phi_i^{k+1}$을 적용하고, (i_k, j_{k+m})이 매칭되지 않으면 $k+1$번째 프레임에서의 평균 위상차 (Δ_Φ)를 적용한다. 따라서 단위음 연결부분에서 매칭되는 정현파는 birth-and-death 법칙에 따라 소멸될 때까지 식 (9)-(11)에서와 같이 연결 합성음에서 그 영향을 미치게 되므로 파형의 모양이 완만히 변화한다.

4.2 다프레임간의 정현파 크기 보간

기존 단위음 연결 방법에서는 단위음간 정현파 크기의 보간을 위해 창함수를 이용하여 중첩가산하거나[3], 식 (12)와 같이 인접한 프레임 간에 크기를 선형보간한다[1].

$$\hat{A}(t) = A_k \cdot W(t) + A_{k+1} \cdot (1 - W(t)) \text{where, } W(t) = 1 - \frac{t}{T} \tag{12}$$

여기서 T는 인접한 두 프레임의 중심점 사이의 거리이다. 그러나 이 방법은 두 단위음의 크기를 두 프레임 사이에서만 보간하므로 음성의 크기가 급격히 변하게 된다. 본 논문에서는 선, 후행 단위음의 여러 프레임을 통해 두 단위음의 크기를 보간하는 방법을 제안한다. 합성음에서 선행 단위음의 최종 프레임을 k번째 합성 프레임이라고 하면, 선행 단위음의 마지막 N_f개의 프레임을 구성하는 정현파들은 $k+1$번째 합성 프레임을 구성하는 정현파와 N_f개의 프레임에 걸쳐 식 (13)과 같이 보간한다. 그리고 후행 단위음의 처음 N_f개의 프레임을 구성하는 정현파들은 k번째 합성 프레임과 N_f개의 프레임에 걸쳐 식 (14)와 같이 보간한다.

$$\hat{A}_{k-i} = W(N+f-i)A_{k-i} + W(i)A_{k+1} \tag{13}$$

$$\hat{A}_{k+i} = W(i)A_k + W(N_f - i)A_{k+i} \tag{14}$$

$$W(i) = \frac{i}{N_f} \text{ , where } 0 \leq i \leq N_f \tag{15}$$

식 (13), (14)에서 $W(i)$는 선, 후행 단위음의 연결프레임에서의 거리에 따른 가중치이다. 그리고 연결부분의 인접한 두 프레임은 단위음 내의 다른 프레임과 마찬가지로 중첩가산한다.
본 논문에서 제안한 방법으로 합성음을 구하면 위상계승을 통하여 파형의 모양이 부드럽게 이어지고, 여러 프레임에 걸친 보간으로 파형의 크기가 완만히 변하는 합성음을 얻을 수 있다.

5 실험 및 결과

본 논문에서 제안한 위상 계승과 다프레임간의 정현파 크기의 보간, 그리고 성도모델의 위상 정보만을 분리해 음성 전체의 위상으로 사용하는 방법을 바탕으로 문서-음성 변환 시스템을 구현하였다. 20ms의 길이의 프레임을 10ms씩 이동시켜 단위음을 분석 하였고, 합성 프레임은 단위음의 지속시간 조절에 따른 가변길이의 합성 프레임을 사용하였다. 분석은 SEEVOC 알고리즘을 사용하여 파워 스펙트럼 상에서 피크를 찾아 각 정현파의 주파수와 위상, 크기 및 스펙트럼 궤적을 구하였다. 사용한 단위음 데이터 베이스는 기존 TTS시스템에[13] 사용되는, VCV(Vowel-Consonant-Vowel) 연쇄음을 기반으로 한 단위음 데이터베이스를 사용하였다. 이 단위음 데이터베이스는 전문 여성 아나운서가 발성하고, 16kHz로 샘플링 된 음성에서 추출되었다. 그림 1은 본 논문에서 제안한 방법으로 구현한 문서-음성 변환 시스템의 합성음이고, 그림 2는 그림 1의 일부를 확대한 것이다. 파형을 살펴보면 파형의 모양이 부드럽게 연결되고, 합성음의 크기가 완만히 변함을 알 수 있다.

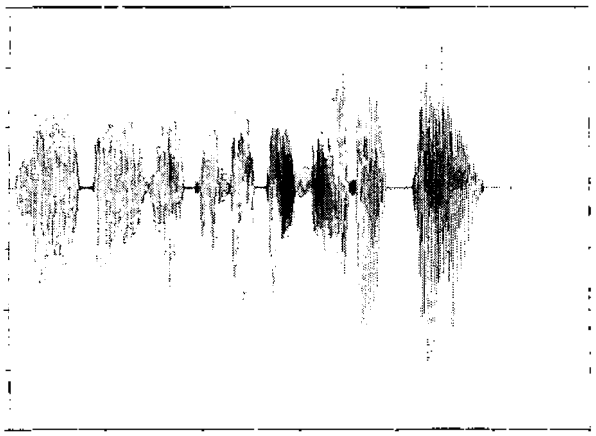

그림 1: "늙은 개구리가 말했다."의 합성음

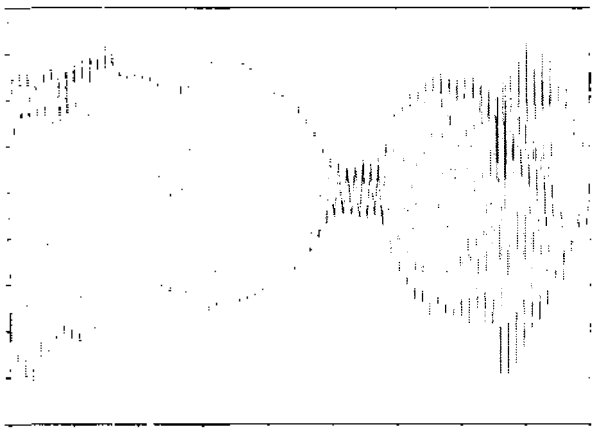

그림 2: 그림 1의 일부

6 결론

본 논문에서는 ABS/OLA Sinusoidal 모델에서 사용되는 피치펄스 정렬을 통한 단위음 연결방법에서 합성음파형의 모양과 크기가 급격히 변하는 문제점과, pitch onset time 계산의 오차로 인해 음성이 왜곡되는 문제점을 해결하기 위하여 위상계승과 다프레임간 정현파 크기의 보간, 그리고 음성의 성도모델의 위상만을 분리하여 합성음의 위상으로 사용하는 방법을 제안하였다. 이를 이용하여 문서-음성 변환 시스템을 구현한 결과 자연스럽게 연결된 합성음을 생성할 수 있었다.

참고 문헌

[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] T. F. Quatieri, R. J. McAulay, "Shape Invariant Time-Scale and Pitch Modification of Speech," *IEEE Trans. on Signal Processing*, vol. 40. no. 3 pp. 497-510, Mar. 1992.

[3] 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, Jun. 1992.

[4] 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, Sep. 1997.

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

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

[7] R. J. McAulay, T. F. Quatieri, "Phase modeling and its application to sinusoidal transform coding," *proc. of ICASSP*, pp. 1713-1715, Apr. 1986.

[8] , Michael W. Macon, Mark A. Clements, "An Enhanced ABS/OLA Sinusoidal Model for Waveform Synthesis in TTS," *proc. of Eurospeech*, vol. 5 pp. 2327-2330, Sep. 1999.

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

[10] R. J. McAulay , T. F. Quatieri, "Sinusoidal Coding" in W.B. Kleijn, K. K. Paliwal, *Speech Coding And Synthesis, Elsevier Science*, 1995.

[11] Thomas W.Parson, "Voice and Speech Processing," McGraw-Hill Book Company, 1986.

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

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