켑스트럼 파라미터와 다중대역 여기신호를 사용한 음성 합성 시스템

°
*김기순, *성유나, *이양희
*동덕여자대학교 전자계산학과

A Speech Synthesis System based on Cepstral Parameters and Multiband Excitation Signal

*Gee Soon Kim, *Yoo Na Sung, *Yang Hee Lee
*Dept.of Computer Science, DongDuck Women's Univ.

요 약

본 논문에서는 명료하고 자연스러운 한국어 음성을 생성하기 위하여 다중대역 여기신호를 이용한 음성 합성 시스템을 제안한다. 분석계에서는 켑스트럼 파라미터를 사용하여 유성/무성 판별 스펙트럼을 이용한 유/무성구간 자동판별법을 제안하고, 현재 단순 임펄스와 백색잡음만으로 구성된 음원과 간단한 유성/무성 판별로 구동되어지는 합성음의 음질상의 한계를 개선하기 위하여 합성계에서는 음질개선 방안으로 유성음 구동시 다중대역 여기신호를 도입하여 합성시 이용한다.

제안된 방법에 대한 청취실험을 한 결과, 유성음 부분 특히 잡음이 많이 섞여있는 유성음화 마찰음과 모음의 천이부분 등에서 일반적으로 사용되고 있는 간단한 유성/무성 파라미터를 사용한 합성음에 비하여 다중대역 여기신호를 사용한 합성음의 명료도가 매우 우수함을 확인하였다.

I. 서 론

임의의 문장을 그에 해당하는 음성으로 출력해 주는 규칙합성 시스템은 맨-머신-인터페이스의 기본적인 부분으로서 자동통역전화 등 많은 분야에서 응용가능하며 현재 이에 대한 활발한 연구가 진행 중이다[1]. 이러한 규칙합성 시스템은, 문자계열을 음운, 운율기호계열로 변환하고, 기호계열을 조음-음원 파라미터 계열로 변환하는 언어처리 시스템과, 이 파라미터 계열을 음성신호로 변환하는 음성합성 시스템으로 구성된다[1].

현재 일반적으로 사용되고 있는 음성합성 시스템은 무성음에 대해서는 백색잡음, 유성음의 경우 임펄스를 음원으로 사용하고 있으며, 이러한 음원과 간단한 유성/무성 판별에 의해 생성되는 음성합성 시스템의 합성음은 음질과 명료성에 한계가 있다[2]. 따라서 본 논문에서는 고품질의 합성음을 생성하기 위하여 다중대역 여기방식을 이용한 분석 합성 시스템을 제안한다. 이 시스템에서는 유성 프레임에 쉬여있는 무성구간의 정보를 복원함으로서 합성음성의 명료도를 높이는 음질 개선 방안을 제안한다.

제안된 음성합성 시스템은 분석시 개량켑스트럼 분석법에 의하여 성도 켑스트럼 파라미터를 추출하며, 추출된 켑스트럼 파라미터를 이용한 유/무성 구간판별 스펙트럼을 생성하여 다중대역 여기신호를 추출하기 위한 유/무성 구간정보 파라미터를 생성한다. 합성시에는 유/무성 구간정보를 이용한 다중대역 여기신호를 생성하여 유성음 합성에 이용하는데, 시간축상에서 하모닉스를 발생시켜 생성한 유성구간 신호와 주파수축상에서 잡음을 발생시켜 생성한 무성구간의 신호를 이용하여 다중대역 여기신호를 합성한다.

제 II절에는 켑스트럼 파라미터와 핏치 및 유/무성 구간정보를 추출하는 분석계에 대하여 기술하고 제 III절에는 다중대역 여기신호를 이용한 음성합성 시스템을 구현하기 위하여 추출된 파라미터로부터 음성을 합성하는 합성계에 대하여 기술하였다. 제 IV절에는 생성된 합성음에 대한 파형을 비교·분석하고 제 V절에는 결론을 맺었다.

II. 분석합성계

고품질의 음성을 합성하기 위해서는 그 기본이 되는 음성합성부인 분석합성계가 명료한 양질의 합성음을 생성해 줄 수 있어야하며, 음성을 분석하여 특징파라미터를 구하므로 그 파라미터가 음성을 잘 모델링 할 수 있어야 한다. 즉, 요구되는 분석합성계는 다음과 같은 조건을 만족시킬 필요가 있다. 1)원 음성의 기본 주파수와 다른 음원에 대해서도 합성음성의 왜곡이 작을 것, 2)스펙트럼포락 파라미터의 시간적인 보간에 있어서 스펙트럼의 왜곡이 적을 것, 3)음성의 기본 단위인 파라미터를 쉽게 추출할 수 있을 것, 4)음성분석에 있어서 스펙트럼포락이 정확하게 추출될 수 있을 것, 5)스펙트럼포락에 근사한 합성필터는 구조가 간단하고 계수 감도가 낮을 것 등이다[6].

본 연구에서는 이를 위하여 주파수 영역에서 정의된 켑스트럼 파라미터를 특징파라미터로 하였으며 합성계는 음성신호가 zero성분을 갖고 있으므로 음성실현의 가능성이 높은 극영모델인 LMA필터를 사용하였다.

일반적으로 음성의 음원은 유성음 음원(periodic

or voiced sound source)과 Turbulent sound source(hissing sound source), Transient sound source(step-like sound source)로 나누는데[10], 간단한 유/무성 판별에 의한 임펄스와 백색잡음만으로 음원을 구동할 경우 이러한 음원의 명료도가 저하된다. 본 연구에서는 시스템의 성능 향상을 위하여 유/무성 정보와 함께, 개량켑스트럼 파라미터를 이용한 유/무성구간 자동판별 스펙트럼을 생성하여 합성시 다중대역 여기방식에 적용하기 위한 유/무성 구간정보를 파라미터로 추출한다.

2.1 분석계

본 연구에서는 개량켑스트럼 파라미터에 의한 분석계를 구현한다. 다음은 본 연구에서 사용한 음성분석계의 개요이다.

- A/D 변환 : 5 kHz LPF,
  10 kHz 샘플링, 12bit 양자화
- 분석/합성 파라미터
  개량 켑스트럼(Improved Cepstrum)
- 분석
  - 프레임 주기
    10 ms : 남성화자, 5 ms : 여성화자
  - Window 함수
    25.6 ms Blackman window
  - 켑스트럼 차수
    30 : 남성화자, 25 : 여성화자
  - 스펙트럼 포락 추출
    개량 켑스트럼법
  - 유성/무성 판별 : 스펙트럼 포락의
    저주파 부분의 에너지 평균
  - 피치 추출
    Cepstrum peak picking법
- 합성 필터
  LMA 필터

분석계에서 추정하는 음원 파라미터는 성도 켑스트럼 파라미터, 유성/무성 정보, 피치정보, 유성/무성 구간 정보이다. 개량켑스트럼 분석법에 의해 음성을 분석하여 성도 파라미터를 구하고, 분석한 각 프레임에 대한 유성/무성을 판별한 뒤 유성 프레임인 경우 피치를 구하고, 다시 유성 프레임에 대해 다중대역 여기신호를 추출하기 위한 유/무성 구간 정보를 추출한다. 그림1은 본 연구에서 사용한 음성분석계의 블럭도이다.

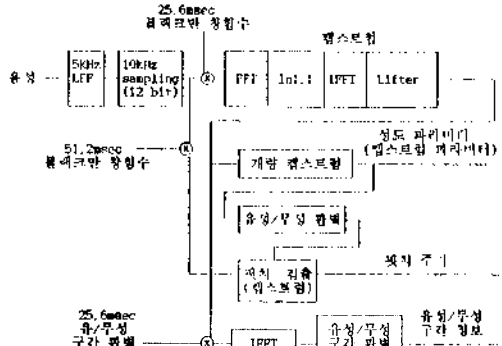

그림1. 음성분석계 블럭도

2.1.1 성도 켑스트럼 파라미터 추정부

켑스트럼은 대수스펙트럼의 푸리에(Fourier)계수로서 정의되지만 음성생성모델의 음원과 조음특성이 각각 음성 켑스트럼의 저차항과 고차항으로 분리되어 나타나는 것을 이용한 분석법[5]이다. 이 분석법에 따르면, 모음과 같은 정상적인 음성신호의 스펙트럼 포락은 인접한 분석 프레임에 있어서 거의 변화하지 않는다. 음성의 대수스펙트럼의 미세구조의 피크치가 안정하기 때문에 이것을 연결하는 유연한 곡선을 스펙트럼 포락이라고 하면, 물리적으로도 의미 있는 좋은 스펙트럼 포락이 된다[6].

음성의 대수스펙트럼 $\ln |S(e^{j\omega})|$은

$$\ln |S(e^{j\omega})| = \ln |H(e^{j\omega})| + \ln |F(e^{j\omega})| \quad (1)$$

$\ln |H(e^{j\omega})|$: 필터의 대수진폭특성
$\ln |F(e^{j\omega})|$: 음원의 대수진폭 특성

로 나타낸다. 켑스트럼은 원식의 푸리에 계수

$$\hat{s}(m) = \hat{h}(m) + \hat{f}(m) \quad (2)$$

로 주어진다. 여기서 $\hat{h}(m)$와 $\hat{f}(m)$은 Quefrency m에 대해 거의 중첩하지 않고 분리할 수 있다. 따라서, 스펙트럼 포락 또는 필터의 특성을 나타내는 켑스트럼 h(m)은 대수스펙트럼 형태로 다음과 같이 나타낸다.

$$\hat{H}(\Omega) = \sum_{m=-M}^{M} \hat{h}(m)e^{-j\Omega} \quad (3)$$

$$= \hat{h}(0) + 2\sum_{m=-M}^{M} \hat{h}(m)\cos(m\Omega)$$

스펙트럼포락 $\hat{H}(\Omega)$은 음성의 대수스펙트럼$\ln|S(e^{j\Omega})|$을 주파수에대해 평활화한 것으로 $\hat{H}(\omega)$을 $\ln|S(e^{j\Omega})|$의 평활화 스펙트럼이라고 한다.

본 연구에서는 스펙트럼의 미세구조에 크게 영향을 받아 변동이 심한 일반적인 켑스트럼보다 log 스펙트럼에 있어서 대략 고조파 리플(미세구조)의 정상을 통과하는 개량 켑스트럼 분석법을 사용한다[5].

2.1.2 유성/무성 판별부 및 피치추정부

유성/무성 판별시, log 스펙트럼 포락의 낮은 주파수 샘플들의 크기를 고려한다. 이러한 샘플들은 유성프레임에서는 제 1 폴만트의 영향을 받아 에너지 레벨이 크게 나타나나, 무성 프레임에서는 저주파수 부분에 폴만트가 존재하지 않아 에너지 레벨이 적다. 유성/무성신호는 스펙트럼 리플의 정상부분을 통과하는 진의 포락의 저주파 영역의 평균레벨 B_v에 의해 자동적으로 결정된다. B_v와 적당한 문턱치와 비교함에 의해, 유성/무성 신호가 결정된다. 남성음성에 대해 약 120 - 400 Hz의 범위에 하중 1, 그 외에서는 0를 부여하여, 매우 좋은

결과를 얻었다. 여성화자의 경우 약 200 - 480 Hz 의 범위에서 구한다[6].

유성 프레임의 핏치 주기는 캡스트럼의 높은 Quefrency 부분에서 핏치의 주기에 대응하는 캡스트럼의 피크치를 추출함으로서 검출된다. 핏치주기 검출기에서 캡스트럼을 사용하기 위해 25.6 msec 의 창함수보다 주파수 분해능이 높은 51.2 msec의 블랙크만 창함수가 사용되며, 유성음인 매 프레임에 대해 3-15msec의 Quefrency영역에서 피크를 찾는다[7].

2.1.3 유/무성구간 자동 판별부

음성 신호는 유/무성음의 혼합된 신호들을 포함하고 있기 때문에 간단한 유성/무성 판별만을 파라미터로 사용하는 일반 합성기에서는 "Buzziness"가 발생하여 음질을 저하시키는 원인이 되고 있다[2]. 본 연구에서는 유성/무성 판별후 유성 프레임에 대한 하모닉스로 구성된 유성 구간과 noise-like energy로 구성된 무성 구간을 결정하여 합성시 다중대역 여기신호를 생성할 수 있도록 한다.

본 연구에서는 유성 프레임의 각 구간에 대한 하모닉스의 유/무성을 판별하기 위하여 캡스트럼의 핏치주기 부근의 High Quefrency 영역을 Liftering 하여 Power Spectrum으로 변환한 결과를 이용한다. 이를 식(4)에 나타낸다.

$$D_{uv}(n) = \frac{1}{N}\sum_{n=0}^{N-1} v_r(n) w_{uv}(n) e^{-j\frac{2\pi}{N}} \qquad (4)$$

(N : 분석 포인트)

$w_{uv}(n)$: 유/무성 구간 판별 창함수

$V_r(k)$: 캡스트럼의 전의 포락

유성/무성 구간 판별을 위해 식(4)로부터 얻어진 Power Spectrum $D_{uv}(n)$의 0~5 kHz 대역을 M개의 채널로 나누어 유/무성 구간판별을 할 때, 각 채널당 5/M kHz(N/2M point)를 유/무성구간 신호 b_{uv}로 나타내게 된다. 유/무성 판별 신호 b_{uv}는 $D_{uv}(n)$에서의 각 채널의 평균치 m_D와 스펙트럼 $D_{uv}(n)$에서의 구간 평균치 d_{uv}, 문턱치 α를 이용하여 결정된다. 문턱치는 실험적 방법에 의해 남성화자인 경우는 0.5, 여성화자인 경우는 0.8이 얻어졌다.

$$b_{uv} = \begin{cases} 1 & \text{유성} \quad (d_{uv} \le \alpha \cdot m_D) \\ 0 & \text{무성} \quad (d_{uv} > \alpha \cdot m_D) \end{cases}$$

그림2는 원음 /아/의 유성 프레임에 대한 0-5 kHz의 스펙트럼(a)과 캡스트럼(b), 유/무성 구간 판별 스펙트럼(c), 유성/무성 구간정보(d)를 나타낸다.

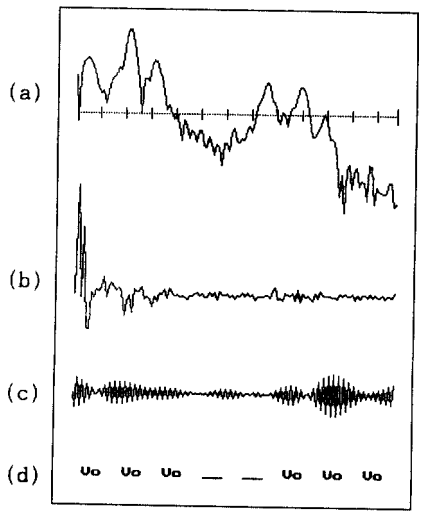

그림2. 유/무성 구간판별 스펙트럼을 이용한 유/무성 구간 정보 추출

2.2 합성계(다중대역 여기신호 생성)

제안한 합성계의 입력은 분석계로부터 보내온 캡스트럼 계수, 핏치 주기 및 유/무성 구간정보이다. 핏치 정보는 유성음 구간에서는 핏치 주기, 무성음 구간에서는 영이 전송되기 때문에 핏치정보가 영일 때는 M-계열을 발생시키는 무성음 합성부, 영이 아닐 때는 다중대역 여기신호를 생성하는 유성음 합성부를 통하여 각 신호를 발생시킨다.

또, 본 연구에서 사용된 시변 디지탈 필터로서는 성도 캡스트럼 파라미터로부터 유도된 캡스트럼 계수에 의해 특징지어지는 Log Magnitude Approximation 필터(LMA 필터, Pole Zero Model)[7]를 합성필터로 사용하고 있다. LMA필터의 log Magnitude 응답은 근사적으로 유한 푸리에 급수이고, 성도전달함수를 나타낸다. LMA필터를 사용한 합성음은 원음과 분석합성음을 비교해 볼 때 분석합성계가 원음을 매우 잘 모델링하여 명료한 합성음을 생성한다[6]. 다음은 본 연구에서 사용한 합성계의 개요를 나타내고 있다.

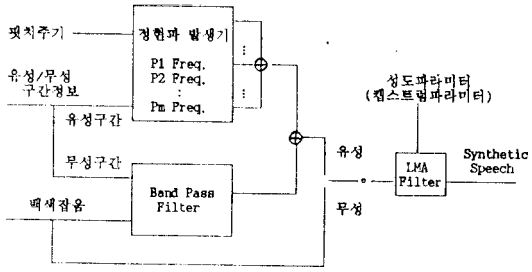

그림3. 음성합성계 블럭도

무성음 합성부에서는 M-계열을 발생시켜 필터의 음원을 만들고, 유성음 합성부에서는 유성/무성 구간정보를 이용하여 다시 유성구간 합성부와 무성구간 합성부로 나뉘어진다. 각 합성부로부터 생성된 유성구간 신호 $s'_{uv}(n)$과 무성구간 신호 $s'_{uv}(n)$을 이용하여 유성 프레임에서의 합성음성 신호를 생성한다.

2.2.1 유성구간 합성부

유성구간 합성부에서는 정현파 발생기를 통하여 프레임내의 유성구간에 대한 각 핏치 정수배 정현파를 발생시키고, 생성된 정현파를 시간영역에서 더하여 유성구간 신호를 구한다[4].

유성구간의 정현파는 핏치주기 정보를 이용하여 유성 프레임내의 기본주파수의 1~ K'배에 해당하는 정현파를 생성한다. 유성구간 신호 $s_v^i(n)$은 다음과 같다.

$$s_v^i(n) = \sum_{k=1}^{K'} \cos(k \cdot \omega_0^i \cdot n + \theta_k^i) \qquad (5)$$
$$(0 \le n \le N)$$

ω_0^i : i번째 프레임의 기본주파수
K' : i번째 프레임의 전체 하모닉스의 갯수
θ_k^i : i번째 프레임의 k번째 정현파 phase
N : 프레임 주기

유성구간 신호는 현(i) 프레임과 이전(i-1) 프레임의 k번째 하모닉스들의 유성/무성 대역에 포함여부를 판별함으로서 결정되어지는데, i번째 i-1번째 하모닉스들이 모두 무성인 경우에는 무성구간 합성부에서 처리되어지고, 그 외의 경우에는 유성인 하모닉스에 한하여 합성 창함수 $w_s(n)$을 이용하여 유성구간 신호를 생성한다. 합성 창함수 $w_s(n)$은 다음과 같다.

$$w_s(n) = \begin{cases} (n+66)\cdot\frac{1}{32} & n < -34 \\ 1 & -34 \le n \le 34 \\ (n-66)\cdot\frac{1}{32} & n > 34 \\ 0 & n > |66| \end{cases}$$

생성된 기본주파수의 k배에 해당하는 각 하모닉스를 시간영역에서 더함으로서 i번째 프레임의 유성 신호를 생성한다.

2.2.2 무성구간 합성부

무성 프레임에 대한 신호를 생성하는 것과는 달리 유성구간 합성부에서의 무성구간 신호는 유/무성 구간정보를 이용하여 백색잡음열 u(n)을 생성하여 무성구간 신호의 합성에 이용한다[3]. 생성된 u(n)과 합성 창함수 $w_s(n)$에 의하여 DFT된 신호 U(n)을 생성하고, 각 채널의 유성/무성 정보에 의해 유성인 경우 0를, 무성인 경우 Noise Spectrum을 이용하여 IDFT를 함으로서 $\hat{U}(n)$을 생성한다.

무성구간 신호 $s_{uv}^i(n)$은 Weighted Overlap Add 방법을 이용하여 다음과 같은 수식으로부터 얻을수 있다.

$$s_{uv}^i(n) = \frac{w_s(n)U^{i-1}(n) + w_s(n-N)U^i(n-N)}{w_s^2(n) + w_s^2(n-N)} \qquad (0 \le n \le N) \quad (6)$$

$w_s(n)$: 합성 창함수
$U^{i-1}(n)$: 이전 프레임에 대한 백색잡음 신호
$U^i(n)$: 현 프레임에 대한 백색잡음 신호

Ⅲ. 실험결과

본 논문에서는 켑스트럼 파라미터와 다중대역 여기신호를 이용하여 유성 프레임에 섞여있는 잡음성분을 합성시 복원함으로서 음성의 명료도를 향상시킬 수 있는 합성시스템을 제안한다. 다중대역 여기신호를 단위음성 합성에 적용하여 실험한 결과 특히 잡음이 많이 섞여있는 유성음과 마찰음과 모음의 천이부분 등에서 간단한 유성/무성 파라미터를 사용한 합성음에 비하여 양질의 합성음을 생성할 수 있었다.

다음은 유성음화 마찰음 /아자/에 대한 전체 파형(그림4)과 부분 파형(그림5)에 대한 것으로, 그림4 그림5의 (a)는 본 합성시스템에 의한 합성음, (b)는 간단한 유성/무성 판별만을 파라미터로 이용한 LMA 합성음, (c)는 원음에 대한 파형이다. 그림4, 그림5에서 다중대역 여기신호를 사용한 (a) 파형이 잡음이 많이 섞여있는 유성음 부분에서 단순 임펄스와 백색잡음만으로 구성된 파형보다 원음과 더 유사함을 볼 수 있다.

본 연구에 의하여 생성된 단위음성에 대한 합

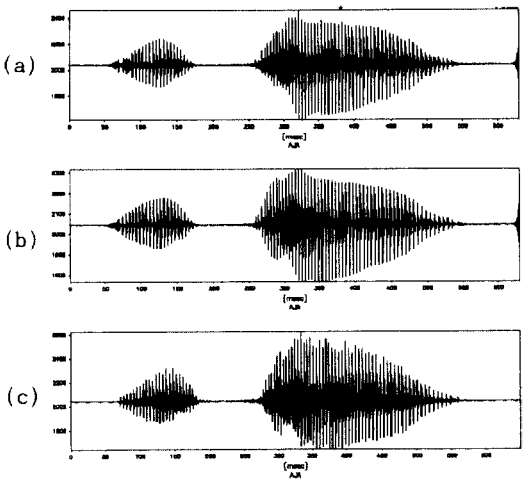

그림4. /아자/에 대한 음성파형

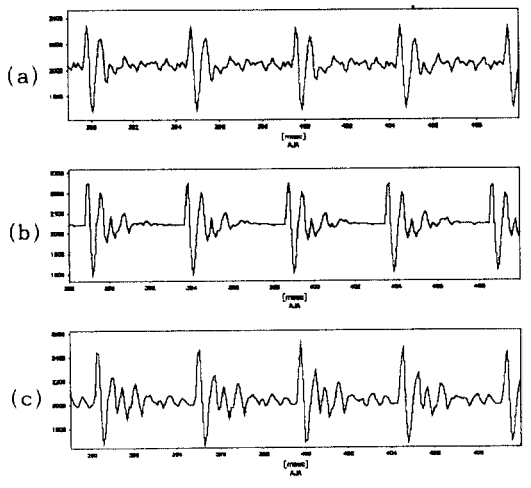

그림5. /아자/에 대한 /자/음의 부분파형

성음을 청취실험을 통한 주관적 평가를 행한 결과 다중대역 여기신호에 의한 합성음의 명료도가 매우 우수함을 알 수 있었다.

Ⅳ. 결 론

본 연구에서는 기존의 단순 임펄스와 백색잡음만으로 구성된 음원과 간단한 유성/무성 판별로 구동되어지는 합성음의 음질상의 한계를 개선하기 위하여 유/무성음이 혼합된 음성 신호들로 구성된 다중대역 여기신호를 생성하는 음성합성 시스템을 제안한다.

이 시스템에서는 유/무성 구간판별 스펙트럼을 생성하여 유/무성 구간정보를 추출하고 시간축상의 하모닉스로 구성된 유성구간 신호와 noise-like energy로 구성된 주파수축상에서의 무성구간 신호를 이용하여 다중대역 여기신호를 생성하고 음성의 명료도를 개선하기 위하여 켑스트럼 파라미터와 함께 합성에 사용한다. 다중대역 여기신호를 단위음성과 문장 합성에 적용하여 실험한 결과 특히 잡음성분이 많이 섞여있는 유성음화 마찰음과 모음의 천이부분 등에서 유/무성음이 혼합된 신호를 합성시 이용함으로서 양질의 합성음을 생성할 수 있었다.

앞으로의 연구 과제로 본 논문에서 제안한 다중대역 여기신호를 이용한 분석합성계를 체계적으로 규칙합성 시스템에 적용한다면 합성음의 명료성과 자연성에 대한 향상을 기대할 수 있으리라 생각된다.

[참고문헌]

[1] 이양희, "음성합성 기술 개발 현황 및 전망", 음성통신 및 신호처리 워크샵, pp.88-93, 1992.

[2] D. W. Griffin and J. S. Lim,"Multiband Excitation Vocoder", Trans. IEEE Vol.36, No.8, Aug. 1988.

[3] J. C. Hardwick and J. S. Lim, "A 4800 bps Improved Multi-Band Excitation Speech Coder", Proc. of IEEE Workshop on Speech Coding for Tele., Sep. 1989.

[4] A. J. Abrantes, J. S. Marques, I. M. Trancoso, "Hybrid Sinusoidal Modeling Of Speech Without Voicing Decision", Proc.Eurospeech, Vol.1, pp.231-234, Sep. 1991.

[5] S. Imai and Y. Abe,"Spectral Envelope Extraction by Improved Cepstral Method." Trans. IECE Vol.62-A No.4, pp.217-223, Apr. 1979.

[6] 정연정, 이양희, "다이폰을 이용한 한국어 음성의 규칙합성 시스템에 관한 연구", 제 10회 음성통신 및 신호처리 워크샵, pp.295-300, 1993.

[7] A. Michael Noll, "Cepstrum pitch determination." J. Acoust. Soc. Am. Vol.41, pp.293-309, Feb. 1967.

[8] S. Imai, et al., "Log Magnitude Approximation (LMA) Filter", Trans.IECE Vol.J63-A No.12, pp.886-893, Dec. 1980.

[9] S. Furui, "Digital Speech Processing, Synthesis, and Recognition", MARCEL DEKKER, pp.168-169, 1992.

[10] J. M. Pickett, "The Sounds of Speech Communication", University Park Press, pp.10-11, 1980.

[11] 지민재, "한국어의 조음 및 음향 음성학", 한국전자통신연구소, pp.47-52, 1993.

[12] 허 웅, "국어음운학", 정음사, 1985.

[13] Y. H. Lee, et al., "A Systematic Conversion Rule for Phonetic Alternants for Use in Korean Speech Synthesis-by-rule.", Trans.IEICE Vol.J71-D No.3 pp.523-530, 1988.

[14] 이양희, "반음절 Mel-cepstrum 파라메터를 사용한 한국어음성의 규칙합성에 관한 연구", 동경공업대학 박사논문, 1988.