포만트 분석/합성 시스템 구현

Implementation of Formant Speech Analysis/Synthesis System

이 준 우* · 손 일 권** · 배 건 성***

(Joon Woo Lee* · Ill Kwon Son** · Keun Sung Bae***)

ABSTRACT

In this study, we will implement a flexible formant analysis and synthesis system. In the analysis part, the two-channel (i.e., speech & EGG signals) approach is investigated for accurate estimation of formant information. The EGG signal is used for extracting exact pitch information that is needed for the pitch synchronous LPC analysis and closed phase LPC analysis. In the synthesis part, Klatt formant synthesizer is modified so that the user can change synthesis parameters arbitarily.

Experimental results demonstrate the superiority of the two-channel analysis method over the one-channel(speech signal only) method in analysis as well as in synthesis. The implemented system is expected to be very helpful for studing the effects of synthesis parameters on the quality of synthetic speech and for the development of Korean text-to-speech(TTS) system with the formant synthesis method.

1. 서 론

음성의 발생원리에 의하면 유성음의 경우 성대(vocal fold)를 통과한 준주기적인 공기의 흐름이 성도(vocal tract)를 지나면서 각 음소에 해당하는 고유한 공명이 성도에서 일어난다. 따라서 유성음의 스펙트럼은 음소마다 고유한 봉우리를 갖게 되는데 이러한 공명봉우리를 포만트(formant)라 하며 낮은 쪽 포만트 주파수에서부터 차례로 제 1, 제 2, 제 3 포만트 등으로 부른다. 무성음의 경우는 불규칙한 공기의 흐름이 성도를 통과하는 동안 성도의 협착점 등에 의한 공명이 발생하게 되며, 따라서 무성음의 스펙트럼에서는 보다 높은 주파수에서 주된 봉우리가 존재하게 된다. 포만트는 음성의 운율적 정보를 결정하는 피치와 함께 음성을 특징 지워주는 중요한 파라미터이며 음성신호의 포만트 주파수와 대역폭을 추정하는 것은 음성인식, 화자

* LG 전자기술원
** LG 전자 TV 설계실
*** 경북대학교 전자 · 전기공학부

인식, 음성 분석 및 합성에 있어서 중요한 사항 중의 하나이다. 이러한 포만트를 음성신호로부터 추정할 때 고려해야 할 사항으로는 1) 음성분석 프레임의 위치와 길이, 2) 포만트와 인접하는 기본주파수의 영향, 3) 비음이나 자음 등에 의해 나타나는 spectral notch, 그리고 4) 자음-모음의 천이구간에서 나타날 수 있는 포만트의 급격한 변화 등이 있다[1].

음성신호의 포만트를 추정하는 방법은 크게 반복적인(iterative) 방법과 비반복적인(non-iterative) 방법, 그리고 수치 해석적인 방법으로 나눠진다[2]. 반복적인 방법은 합성에 의한 분석법(analysis by synthesis)으로 분석한 스펙트럼과 음성발생모델에 의해 합성한 스펙트럼을 비교하여 그 오차정보에 의해 합성을 위한 포만트 주파수 등을 갱신하는 과정을 반복하여 점진적으로 포만트를 추정하는 방법으로, 그 성능에 비하여 반복을 위한 많은 시간이 필요하고 정확한 음성발생모델의 선정이 요구된다는 단점이 있다. 비반복적인 방법은 평활화된 스펙트럼에서 peak picking을 행하는 방법으로 cepstral 스펙트럼 또는 LPC(linear predictive coding) 스펙트럼이 흔히 사용되며, 수치 해석적인 방법은 가장 일반적인 음성발생모델인 선형필 터모델에서 전달함수의 극점을 구하여 포만트를 구하는 것으로 Newton-Raphson 방법이 주로 이용된다[2~7]. 일반적으로 많이 사용되고 있는 포만트 추정방법은 LPC 분석을 통해 선형예측계수를 구한 뒤, LPC 다항식에서 root solving 이나 LPC 스펙트럼에서 peak picking을 행하는 것이다. 이때 음성분석 프레임의
크기나 위치 등이 포만트 추정에 많은 영향을 미칠 수 있게 된다. 이러한 문제점들을 해결하기 위해 pitch synchronous analysis[1] 또는 closed phase analysis[4]를 행하게 되는데 이를 위해서는 음성신호의 보다 정확한 피치정보를 필요로 하게 된다.

이 연구에서는 포만트 합성방식의 문자-음성 변환 시스템을 구현하기 위한 기초연구로, 음성신호의 분석을 통해 정확한 포만트 정보를 추출할 수 있는 알고리즘을 개발하고, 음원파형이 합성음에 미치는 영향을 비롯하여 문자-음성 변환 시스템의 구현에 필요한 다양한 분석 및 합성 실험을 할 수 있는 포만트 분석/합성 시스템을 구현하였다. 포만트 추정 방법으로는 음성만을 이용한 1-채널 방식의 포만트 추정 방법과 함께 주변환경 및 잡음의 영향을 받지 않고 성대의 떨림에 관한 정보를 정확하게 전달해 주는 EGG(electroglottograph) 신호를 이용한 2-채널 음성분석 알고리즘을 구현하여 보다 정확한 포만트 정보를 얻고자 하였다. 2-채널 음성분석은 먼저 EGG 신호의 분석을 통해 유성음의 정확한 피치정보를 얻고 이를 이용하여 유성음/무성음/묵음 분류를 행한 뒤 유성음 구간에서는 피치단위 또는 closed phase 구간 단위로, 무성음 구간에서는 가변적인 프레임단위로 선형예측계수를 구하여 root solving 함으로써 포만트를 추정하였다.

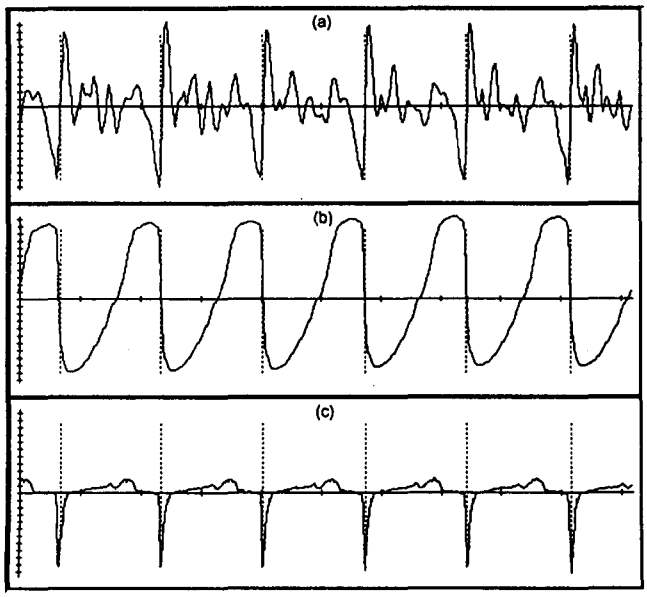
실제 분석/합성 과정에서 여기신호의 파형 및 음성분석에서 얻어지는 합성파라미터들은 합성음의 음질에 많은 영향을 미치게 된다. 따라서 본 연구에서는 여기신호원으로 2-pole 모델, LF 모델과 Rosenberg에 의해 제안된 모델[8,9]을 이용하여 여기신호와 관련된 여러 파라미터들을 변화시키면서 합성음에 나타나는 영향에 대해서도 연구 및 실험하였으며, 다양한 분석 방

법에 의해 얻어지는 포만트 특징 및 파라미터들이 합성음의 음질에 어떤 영향을 미치는가를 연구하였다.

2. 음성 및 EGG 신호를 이용한 음성신호의 분석

EGG 신호는 성대의 진동운동을 간접적으로 관측하기 위해 고안된 방법으로 성대의 떨림을 임피던스 변화로 바꾸어 전기신호로 나타낸 것이다[10,11]. 성대가 서로 떨어져 있을 때는 열린 공간을 통해 전류의 통로가 형성되므로 두 전극간의 임피이던스가 큰 값이 되고, 성대가 서로 붙어있을 때는 얇은 막을 통해 전류의 통로가 형성되므로 임피이던스가 작은 값을 가지게 된다. 따라서 유성음을 발성할 경우 성대의 진동운동에 의해 단조증가 및 단조감소 현상을 나타내는 단순한 형태의 EGG 신호를 얻게 된다. 음성신호와 달리 EGG 신호는 포만트의 영향을 받지 않고 성대진동을 그대로 반영하므로 그림 1에서와 같이 피치주기 검출에 EGG 신호를 미분한 형태인 DEGG(differentiated EGG) 신호의 부(-)의 최대 피크값들 사이의 간격을 계산함으로써 피치주기가 얻어지므로 매우 간단하고도 정확한 결과를 얻을 수 있다. 본 연구에서는 2-채널 방식의 포만트 추정에 필요한 피치정보 및 유성음/무성음/묵음의 분류정보를 얻기 위해 음성 및 EGG 신호 분석에 의한 피치검출 알고리즘을 수정하여 사용하였다[12].

그림 1. (a) 음성파형 (b) EGG 신호 (c) DEGG



가장 일반적으로 사용되고 있는 포만트 추정방법은 음성발생모델에서의 성도의 특성을 식 (1)과 같은 all-pole 필터로 표현하는 LPC 분석에서 시작된다.

$$H(z) = \frac{1}{1 + \sum_{k=1}^{P} \alpha_k \ z^{-k}} \tag{1}$$

여기서, p는 LPC 차수이고 α_k(k = 1, 2, · · p)는 LPC 계수이다. 식 (1)로 주어지는 성도의 전달함수 H(z)의 스펙트럼은 음성신호 스펙트럼의 포락선을 나타내고 포만트에 해당하는 위치에서 피크치를 가진다. 따라서 이러한 스펙트럼에 대한 peak picking을 행하거나 극점들을 직접 계산하는 방법으로 포만트를 구하게 된다. 본 연구에서는 전달함수 H(z)의 complex conjugate pole로 나타나는 포만트를 수치 해석적 root solving 방법으로 구하였다. 특히 유성음 구간에서는 그림 2와 3에 보인 것과 같이 분석프레임을 해당 피치의 시작점과 끝점으로 하는 피치 동기적인 방식과 성문의 폐쇄구간을 분석프레임으로 하는 closed phase 분석을 행하여 공분산방법으로 LPC 계수를 구하였으며, 무성음 구간에서는 스펙트럼 특성 변화에 따른 가변적인 분석프레임에 대해 자기상관방법으로 LPC 계수를 구하였다. 각 방법에 따른 분석조건은 다음과 같다.

- 고정프레임 분석(fixed-frame analysis)
  분석프레임의 길이는 200 샘플(10 kHz 표본화주파수에서 20 msec)로 하고 100샘플씩 중첩시켜 Hamming 창함수를 곱한 뒤, 자기상관방법을 사용하여 선형예측계수를 구하였다.

- 피치동기식 분석(pitch synchronous analysis)
  분석프레임의 길이는 그림 2에서와 같이 피치주기(T)와 LPC 분석과정에서의 초기조건으로 사용하기 위한 피치 시작점전의 p 샘플을 합한 것으로 하였고, LPC 분석은 공분산방법을 사용하였다. 여기서 p는 LPC 분석의 차수이다.

- closed phase 분석
  분석프레임의 길이는 그림 3에서와 같이 성문이 닫히는 시점에서 열리는 시점까지인 성문폐쇄구간 (C)에 LPC 분석과정에서의 초기조건으로 사용하기 위한 p 샘플을 합한 것으로 하였고, LPC 분석에는 공분산방법을 사용하였다.

그림 2. 피치동기식 분석프레임

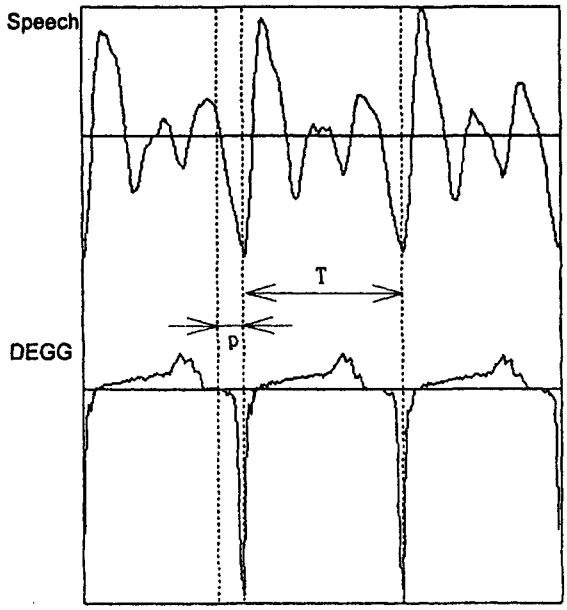
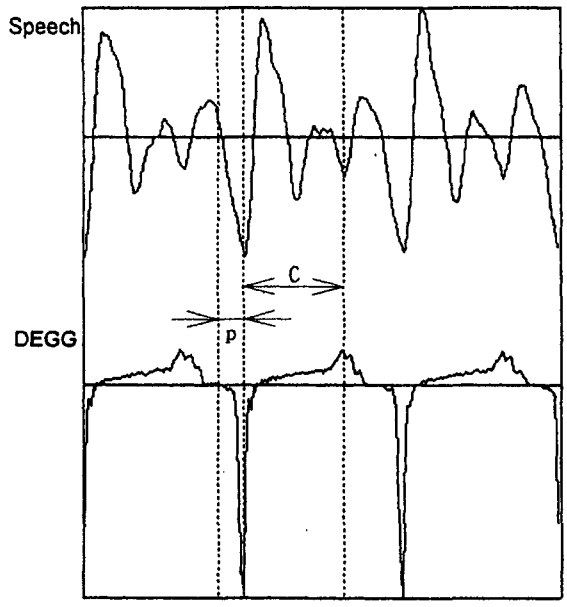


그림 3. closed phase 분석프레임



유성음의 경우와는 다르게 무성음은 낮은 주파수 영역에 큰 대역폭을 가지는 포만트가 존재하는 경우가 많은데, 안정된 극점만을 포만트로 받아들이기 위해 복소평면의 단위원에 근

접한 극점만을 포만트로 간주할 경우 낮은 주파수 영역의 광대역 포만트는 추정되기가 어렵다. 따라서 본 연구에서는 성도모델에 존재할 수 있는 음향학적인 공진주파수의 개수(5 kHz까지 4-5개)를[13,14] 만족시킬 수 있도록 복소평면상의 포만트 추정 범위를 점진적으로 넓힘으로써 무성음 또는 일부 유성음의 낮은 주파수 영역에 존재하는 포만트 추정의 정확도를 높였다. 또한 유성음의 경우도 무성음의 경우와 마찬가지로 3 kHz 이상의 주파수 영역에서는 이론적인 개수보다 많은 공진주파수가 존재할 수 있어 고정된 차수의 LPC 분석으로는 정확한 포만트의 추정이 어렵게 된다. 따라서 포만트 추정범위의 변화와 더불어 LPC 분석 차수를 변화시킴으로써 보다 정확한 포만트 추정치를 얻고자 하였다. 제안된 포만트 주파수 및 대역폭 추정 알고리즘은 다음과 같다.

i) LPC 분석을 통해 선형예측계수를 구하여 p차의 LPC 다항식을 구한다.

ii) LPC 다항식으로부터 complex conjugate pole(z_i)을 구한다.

$$z_i = r_i e^{j\theta_i}, \quad i = 1, 2, \ldots$$

iii) 복소평면상에서 단위원 밖에 있는 pole($|z_i| > 1$)의 역수를 취해 단위원 안으로 이동시킨다.

$$z_i = \frac{1}{r_i} e^{j\theta_i}, \quad i = 1, 2, \ldots$$

iv) 대역폭이 비정상으로 넓은($B > 500$ Hz) 포만트를 제거하기 위해 $|1 - r_i|$ 값이 작은 값을 갖는 순서대로 구하고 싶은 포만트 수만큼 선택한다.

v) 구해진 pole의 개수가 음향학적인 개수를 만족시키지 못할 경우 대역폭의 임계치를 800 Hz까지 100 Hz 씩 증가시켜 iv)의 단계를 반복한다.

vi) 대역폭의 임계치를 800 Hz까지 증가시켜도 원하는 수만큼의 pole을 구하지 못할 경우 LPC 차수 p를 16차까지 2 만큼씩 증가시켜 i)의 단계로 되돌아간다.

vii) 아래의 식들을 이용하여 포만트 주파수 F_i 및 대역폭 B_i 를 계산한다. 여기서 T는 샘플링주기이다.

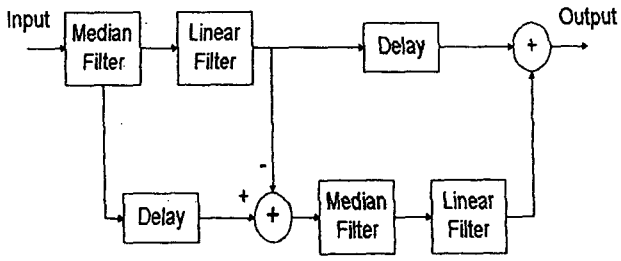
$$F_i = \frac{\theta_i}{2\pi T} \tag{2}$$

$$B_i = \frac{1}{\pi T} \ln \frac{1}{r_i} \tag{3}$$

추정된 포만트 주파수와 대역폭은 프레임 간에 불연속성을 보일 수 있다. 이러한 불연속성은 음성의 발생구조상으로는 일어날 수 없는 것이므로 적절한 평활화 과정이 필요하게 된다. 평활화 알고리즘은 크게 선형 평활화와 median 필터로 대표되는 비선형 평활화로 나눠지는

데 각각을 단독으로 사용할 경우 원래의 데이타를 지나치게 감쇄시키는 효과를 가져오게 되므로 좋은 방법이라 할 수는 없다. 본 연구에서는 그림 4와 같이 선형 평활화와 median 필터를 복합적으로 사용한 비선형 평활화 알고리즘을 적용하였다[15].

그림 4. 평활화 알고리즘 도식



3. 포만트 음성합성

포만트 음성합성기는 각각의 포만트 주파수 및 대역폭 특성을 갖는 디지탈 공진기를 구성하고 이를 적절한 음원으로 여기(excitation)시켜 음성을 합성하는 방식으로 간단한 시스템의 경우 대역폭은 고정시키고 처음 3개만의 포만트를 변화시켜 합성을 수행하고 좀 더 진보된 시스템의 경우 포만트 주파수와 대역폭을 모두 변화시켜 합성음을 얻는다. 포만트 합성기는 디지탈 공진기 즉, 디지탈 필터를 구성하는 방법에 따라 직렬합성기와 병렬합성기, 그리고 두 합성기가 가지는 장점을 살린 직·병렬합성기로 나눠진다. 직렬합성기는 디지탈 공진기의 출력이 다음 공진기단의 입력으로 이용되는 구조이며 합성시 각 포만트에 대한 이득을 일일이 조정하지 않아도 된다는 점과 그 구조상 비음 특성을 갖지 않는 모음(non-nasal vowel)의 합성에 적절하다는 점을 장점으로 들 수 있으나, 자음의 합성을 위해 복잡한 추가 회로가 필요하다는 단점이 있다. 반면에 병렬합성기는 마찰음, 파열음 등 자음이 가지는 다양한 스펙트럼 특성을 효율적으로 조정할 수 있다는 장점이 있으나 각 포만트에 대한 이득이 미리 구해져야 하며 모음의 합성시 낮은 주파수영역에 존재하는 포만트와 기본 주파수와의 영향이 가져오는 합성음질의 저하를 피하기 어렵다는 단점이 있다[16,17]. 각각의 합성기가 지니는 이러한 장·단점을 고려하여 그들의 장점만을 살리고자 Klatt는 일반적인 모음의 합성은 직렬합성기가 전담하게 하고, 마찰음, 파열음 등 자음의 합성은 병렬합성기, 또는 경우에 따라서 두 합성기를 동시에 사용하는 직·병렬합성기를 구현하여 양호한 합성음질을 얻은 바 있다[18].

이 연구에서는 Klatt의 직·병렬합성기를 구현하여 음성분석의 결과로 얻어지는 피치정보, 유성음/무성음/묵음 분류정보 그리고 포만트 정보 등을 이용한 합성 실험을 하였으며 고정프레임방식, 피치동기식, 그리고 closed phase 방식 등과 같은 포만트를 추정하는 방법에 따른 분석 및 합성결과를 비교하였다. 또한 합성파라미터의 변화가 합성음질 및 파형에 미치는 영

향을 관찰할 수 있도록 합성파라미터와 음성파형 및 음성의 스펙트로그램을 화면에 나타내고 마우스를 사용하여 파라미터 값을 사용자가 적절히 변경하여 합성할 수 있는 합성파라미터 display & modify tool을 구현하였다.

공기가 좁은 성문을 통과해서 흐를 때 성문에 걸리는 압력은 성문 전후 공간의 크기에 의한 음향적 loading 효과에 의해서 매 진동 주기 동안에 변하게 되므로 glottal 펄스 파형의 정확한 모델링은 매우 어렵다. 때문에 1차원 flow 이론과 성도의 진동을 표현하기 위한 2-mass 모델, 그리고 flow와 성문의 면적 함수와의 관계를 나타내기 위한 등가면적을 사용하여 일반적으로 모델을 단순화한다. 음성발생시 glottal 펄스 파형은 사람마다 다르며, 양질의 음성합성을 위해서는 성도뿐만 아니라 음원의 특성을 설명할 수 있는 수학적 모델링이 중요하므로 실제 glottal 펄스 파형과 비슷한 형태의 여기신호를 만들기 위한 많은 연구가 있었다[19].

2-pole 모델은 피치 주기마다 임펄스를 여과시켜 2-pole로 구성된 필터를 거쳐 나온 신호를 여기신호로 사용한다[16]. 이때 사용되는 필터는 -12 dB/octave의 특성을 가진다. LF 모델은 Fant, Liljencrants와 Lin에 의하여 제안되었으며, 적은 수의 파라미터로 일반적으로 보게 되는 glottal 펄스의 모양에 전체적으로 맞도록 독립적인 파라미터들을 구하는데 적당하다[9]. 합성기에 입력되는 미분된 glottal 펄스의 형태는 식 (4)로 주어진다.

$$E(t) = E_o e^{at} \sin(w_g t) \qquad 0 < t < t_e \qquad (4.a)$$

$$E(t) = -E_e / \zeta t_a [e^{-\zeta(t-t_e)} - e^{-\zeta(t-t_c)}] \qquad t_e < t \leq t_c \qquad (4.b)$$

Rosenberg 모델은 glottal 펄스의 진폭, 지속시간과 기울기의 변수들을 사용해서 2개의 trigonometric segment들에 의해 식 (5)로 모델링된다[8].

$$U_g(t) = k_1 t^2 - k_2 t^3 \qquad (5.a)$$

$$\frac{dU_g(t)}{dt} = \text{sum over t = 1, 2, ..., } n[a-(bt)]$$

$$= an - \frac{b}{2} n^2 \qquad (5.b)$$

무성음의 여기신호는 동요가 큰 잡음여기원을 사용하는데, 합성기에서는 난수(random number)발생기로 발생시킨다. 이때의 스펙트럼은 -6 dB/octave의 주파수 특성을 갖도록 이 여기원의 volume velocity는 여기원 압력의 적분의 형태로 주어지며, 이 때의 적분은 차수가 일차인 low-pass filter로 근사화 된다.

4. 실험 및 검토

이 연구에서 음성의 포만트 분석 및 합성 실험에 사용된 데이터는 남성화자에 의해 녹음된 영어와 한국어 문장 각 1 개씩의 음성신호 및 그에 동기된 EGG 신호이다. 문장의 내용과 분석조건은 다음과 같다.

영어문장 : "We were away a year ago."
한국어문장 : "힘이 더 세다고 하기로 결정했습니다."
Sampling rate : 10 kHz, 16 bits/sample
Preemphasis factor : 0.95

여기신호의 음원파형에 따른 합성음의 음질변화를 실험하기 위해 사용된 문장 및 분석조건은 다음과 같다.

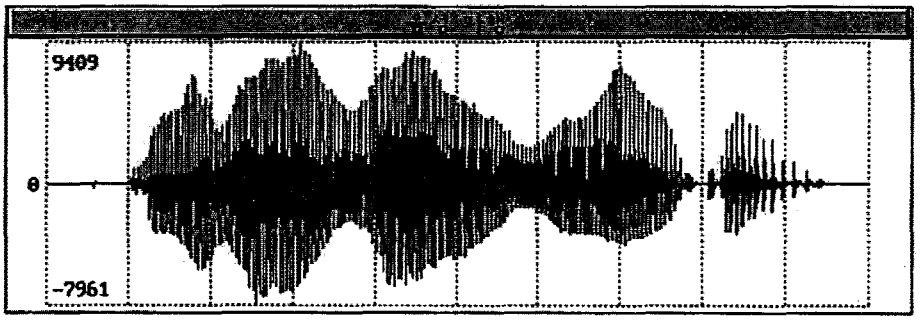
영어문장 : "We were away a year ago."
"Should we chase those cowboys?"
Sampling rate : 10 kHz, 16 bits/sample
Preemphasis factor : 0.95

음성의 분석과정에서 포만트 주파수와 대역폭은 앞에서 설명한 바와 같이 대역폭 임계치의 조정과 LPC 차수를 변화시켜 root solving 함으로써 낮은 주파수 영역의 광대역 포만트와 높은 주파수 영역의 포만트의 추정 정확도를 높였다. 음성신호만을 이용한 1-채널 분석인 고정 프레임 분석결과와 2-채널의 피치동기식, closed phase 분석결과를 비교하여 2-채널 음성분석이 1-채널 음성분석기 성능의 검증 척도가 될 수 있음을 확인하였다. 영어문장 "We were away a year ago"를 각각의 방식으로 분석하여 얻어진 포만트 주파수 및 대역폭, 기본주파수 궤적 그리고 에너지를 그림 5에 나타내었다. 고정프레임 분석에 의해 구해진 포만트 궤적은 그림 5(b)에서 보듯이 /ə I/부분에서 포만트의 급격한 변화를 따라가지 못하여 두 번째, 세 번째 포만트에서 잘못된 추정이 많이 이루어짐을 볼 수 있고 그에 비해 피치동기식과 closed phase 분석에 의해 구해진 포만트 궤적은 매우 안정된 포만트궤적을 얻을 수 있었다. 한국어 문장의 분석에서는 특히 유성음/무성음의 천이 구간에서 2-채널의 분석결과가 1-채널 분석결과에 비해 훨씬 뚜렷한 포만트 변화를 보여줌을 알 수 있다.

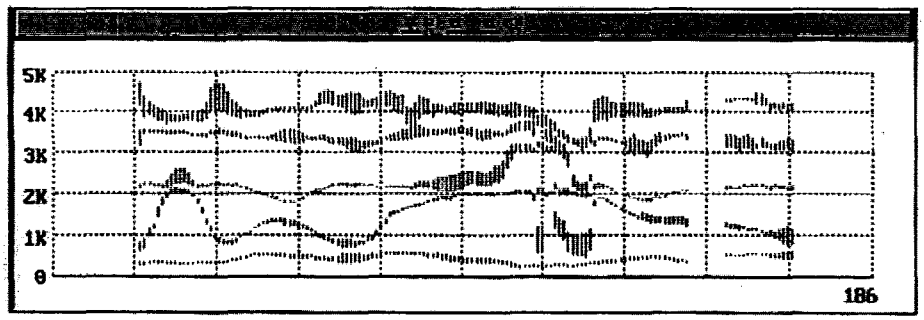
분석에 의한 음성합성 시스템의 성능은 얼마나 원음에 가까우면서도 자연스러운 합성음을 만들 수 있는가에 달려 있다. 원음과 합성음을 비교하는 방법으로는 청취 실험을 통한 주관적인 방법과 시간영역 및 주파수 영역에서의 특성을 비교하는 방법이 있는데 시간영역의 비교는 파형의 모양을 통해서, 주파수 영역의 비교는 스펙트로그램을 통해서 이루어 질 수 있다. 그림 6은 closed phase 방식에 의해 구해진 합성 파라미터들을 이용하여 합성된 영어문장

의 합성 파형과 원음의 주파수특성을 비교한 것이다. 합성에 사용된 포만트 파라미터 값이 원음의 주파수 특성을 잘 반영함을 볼 수 있으며, 청취 실험의 결과 또한 원음에 가까운 양호한 음질을 가짐을 확인하였다. 그리고 각 분석방법에 따른 합성음의 음질은 2-채널 분석에 의한 합성음이 1-채널에 의한 것보다 명료성에 있어서 뛰어난 것으로 나타났는데, 이는 1-채널 분석이 포만트 변화를 제대로 따라가지 못했기 때문인 것으로 보여진다. 합성을 위한 여기원으로 무성음의 경우에는 백색잡음을, 유성음의 경우엔 Rosenberg 모델을 사용하였다.

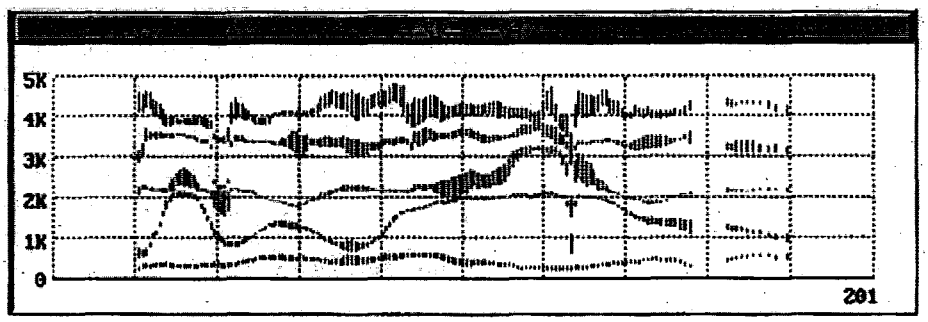
그림 5. 영어문장 "We were away a year ago"의 분석결과



(a) Original speech



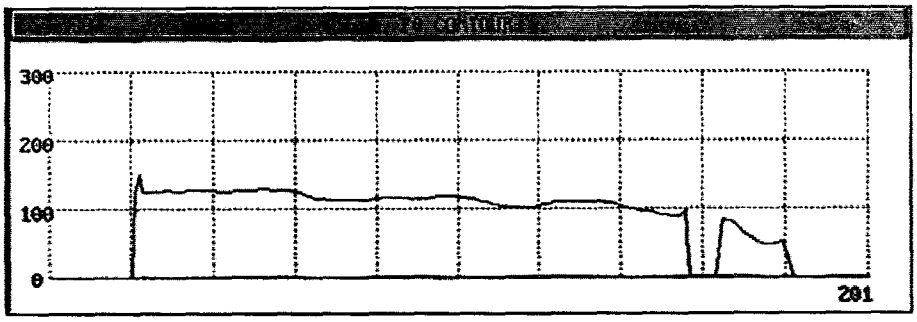
(b) Formant track using pitch asynchronous analysis



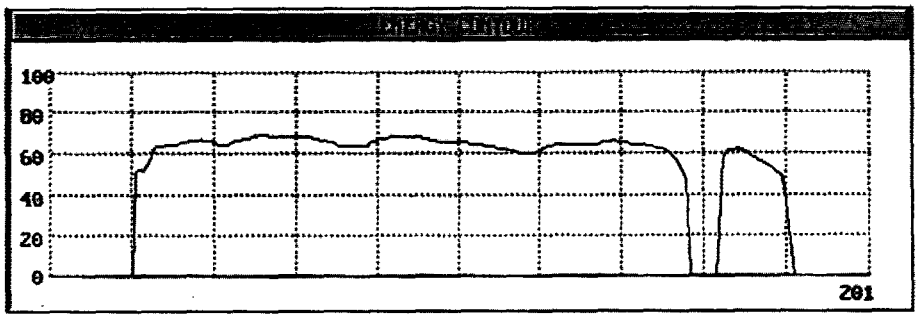
(c) Formant track using pitch synchronous analysis

그림 5. (계속)

(d) Formant track using closed phase analysis



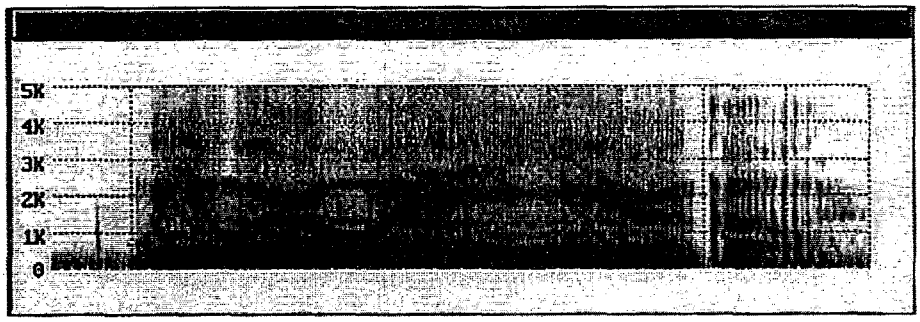
(e) Pitch contour



(f) Energy contour

그림 6. 합성된 음성신호 "We were away a year ago."

(a) Synthetic speech



(b) Spectrogram of the

(c) Formant track

여기신호에 따른 합성 실험에서, 포만트 합성에 필요한 파라미터들은 2-채널(음성 및 EGG) 신호분석[8]에 의해 얻어진 값들을 이용하였다. 포만트 주파수 및 대역폭은 2-채널 신호를 사용하여 유성음/무성음/묵음 분류를 한뒤 피치동기방식에서는 covariance 방법으로, 피치비동기방식에서는 autocorrelation 방법으로 구해졌으며, 합성형태는 피치동기방식을 사용하였다. 그리고 glottal 펄스로 사용되는 각 여기신호의 모델에서 open quotient와 관련되는 각 파라미터를 표 1에 나타내었다. 실제적으로 음성에서 적당한 open quotient를 찾아서 적용하여야 하나, 이 실험에서는 전체 합성단위에 일정한 비율로 open quotient를 30 %, 50 %, 70 %와 90 %로

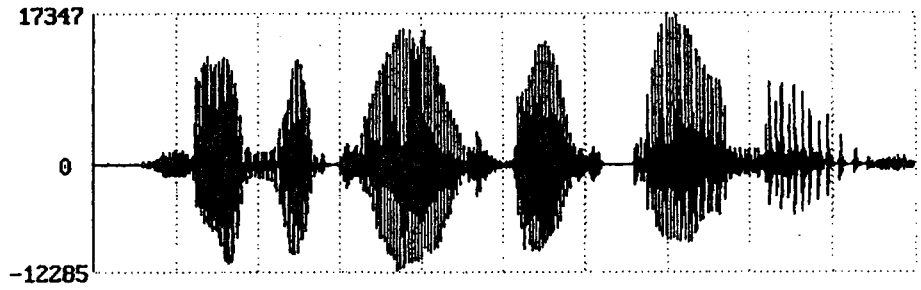
변화시키면서 합성을 하였다. 그림 7은 남성화자에 의해 발성된 문장 "Should we chase those cowboys?"의 음성신호와 피치동기방식으로 분석된 파라미터들을 보인 것이다. 이를 이용하여 여러 모델의 여기신호에 따른 합성음을 그림 8에 나타내었다.

표 1. 여기신호 모델의 파라미터 값

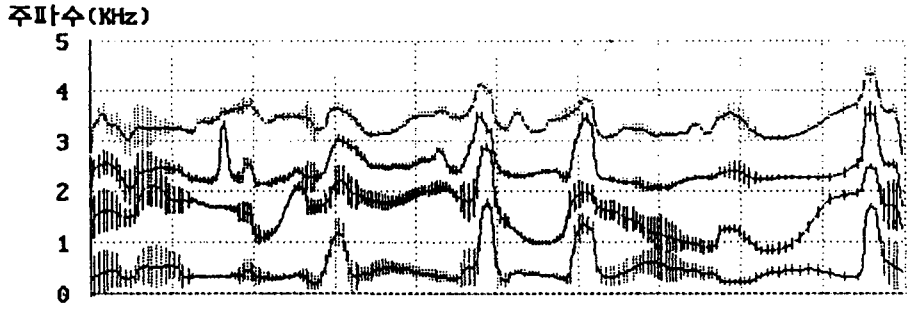
여기신호	파라미터
2-pole 모델	공진기의 입력값 대역폭 = sampling rate / open period
LF 모델	timing 파라미터 t_o = 피치주기 t_c = open quotient * 피치주기 $t_e = t_c \times 0.1$ $t_p = t_c \times 0.6$ $t_a = t_c \times 0.8$ E_e = gain
Rosenberg 모델	trigonometric 상수 : b = gain / (open period)^2 $a = b \times$ (open period / 3)

그림 9는 그림 7의 영어문장에 대해 Rosenberg 모델을 이용하여 glottal open quotient를 변화시켜 가면서 합성된 결과이다. 전체적인 파형의 형태에는 큰 변화가 없음을 볼 수 있다. 여러 실험의 결과를 살펴보면 각 여기신호에 대한 합성음은 LF 모델의 명료성이 다른 여기신호보다 뚜렷이 나타났고, 원음과의 유사성은 Rosenberg 모델이 보다 가까웠다. 그리고 각 여기신호의 open quotient를 적게 줄수록 명료성은 잘 나타나고, 자연성은 open quotient를 크게 할수록 잘 나타났다. 따라서 각 여기신호에 따라 open quotient를 변화하면 다양한 합성음을 얻을 수 있다. 그림 10은 본 연구에서 개발한 포만트 분석/합성 시스템의 합성파라미터 display & modify tool을 나타낸 것으로 포만트 주파수 및 대역폭, 에너지, 그리고 기본주파수를 표시할 수 있는 표시모드와 사용자가 지정한 합성파라미터를 마우스를 이용하여 임의로 수정·합성할 수 있는 갱신·합성모드로 구성되어 있다. 개발된 도구를 이용하여 음성신호의 분석된 결과에서 사용자가 임의로 파라미터를 변화시켜 가면서 합성파라미터가 합성음에 미치는 영향에 관한 연구를 interactive하게 수행할 수 있다.

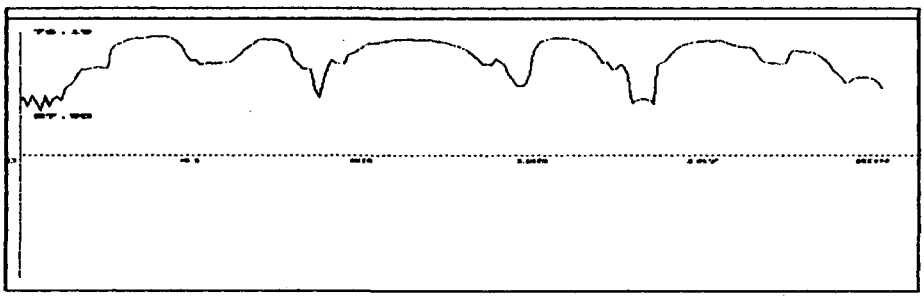
그림 7. 음성신호 및 분석결과 ("Should we chase those cowboys?")



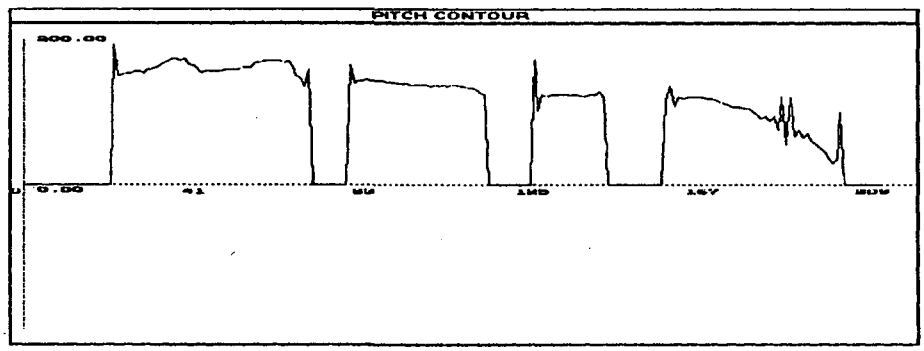
(a) Original speech



(b) Fromant track

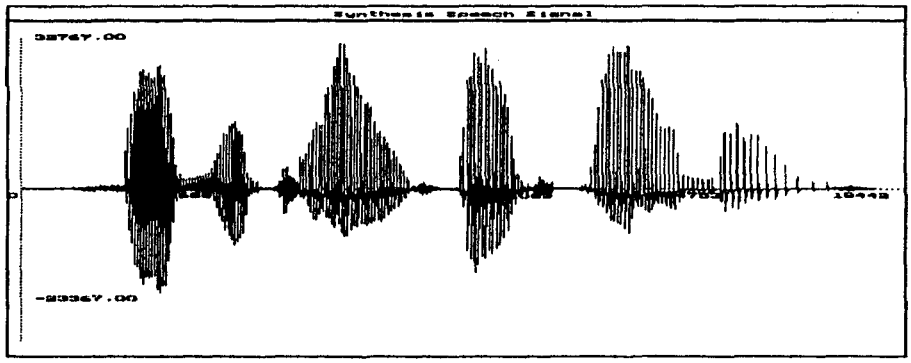


(c) Pitch contour

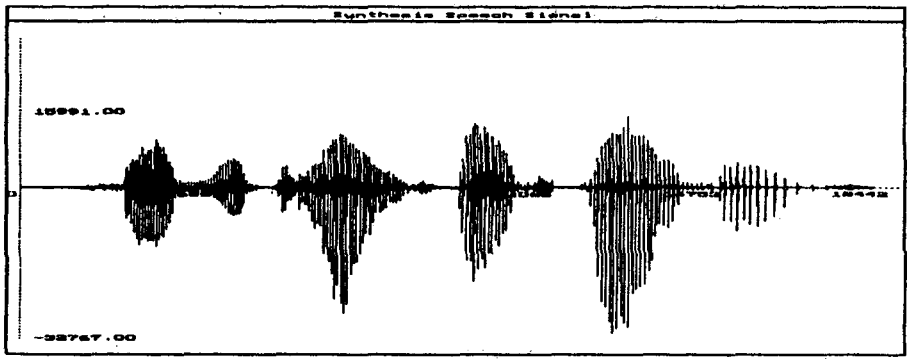


(d) Energy contour

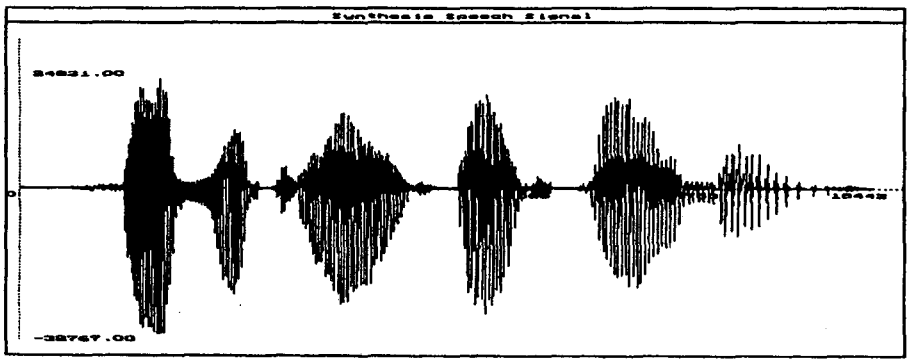
그림 8. 여기신호에 따른 합성음



(a) Synthetic speech with 2-pole model

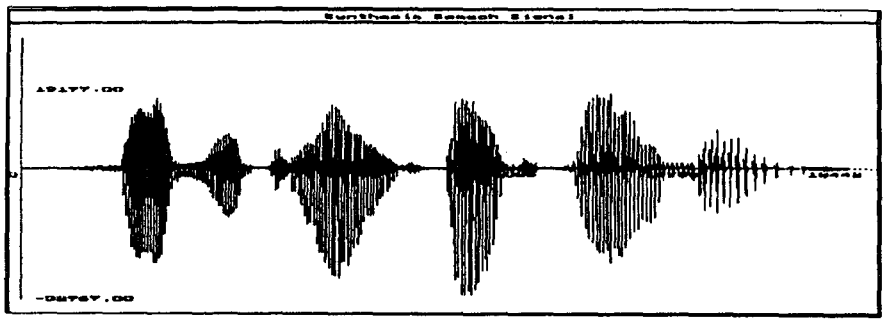


(b) Synthetic speech with LF model

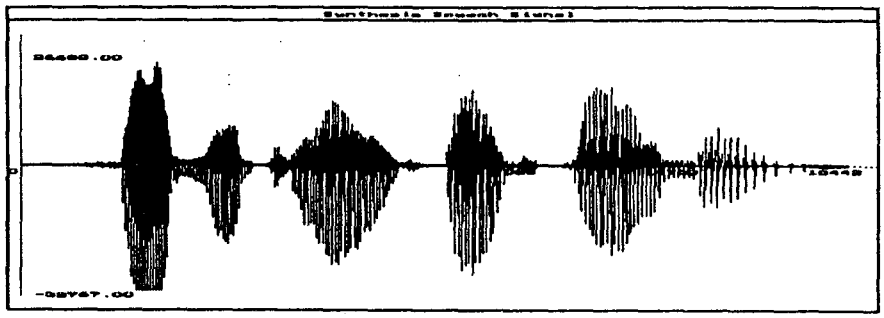


(c) Synthetic speech with Rosenberg model

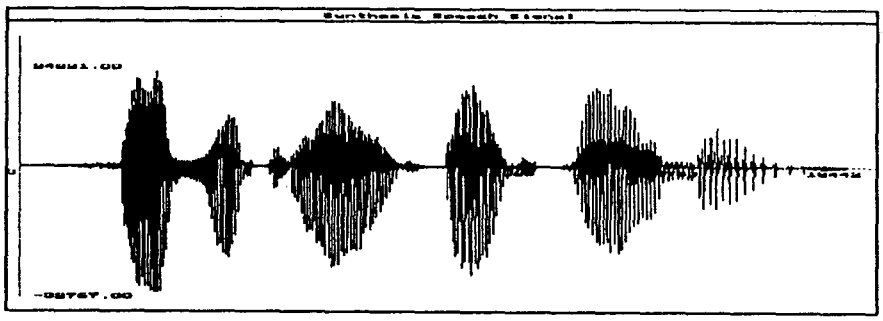
그림 9. Open quotient 변화에 따른 합성음



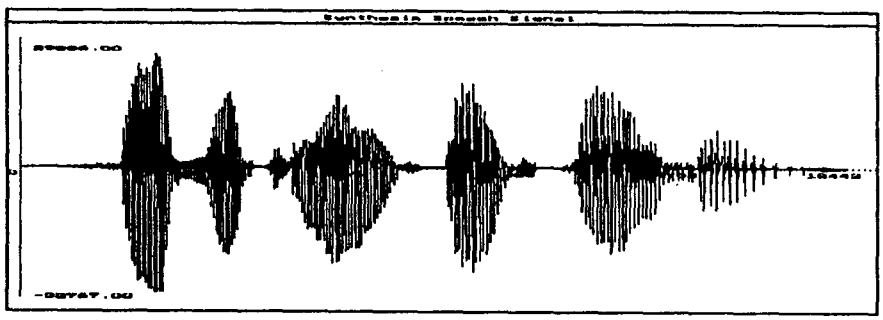
(a) Open quotitent 30%



(b) Open quotient 50%

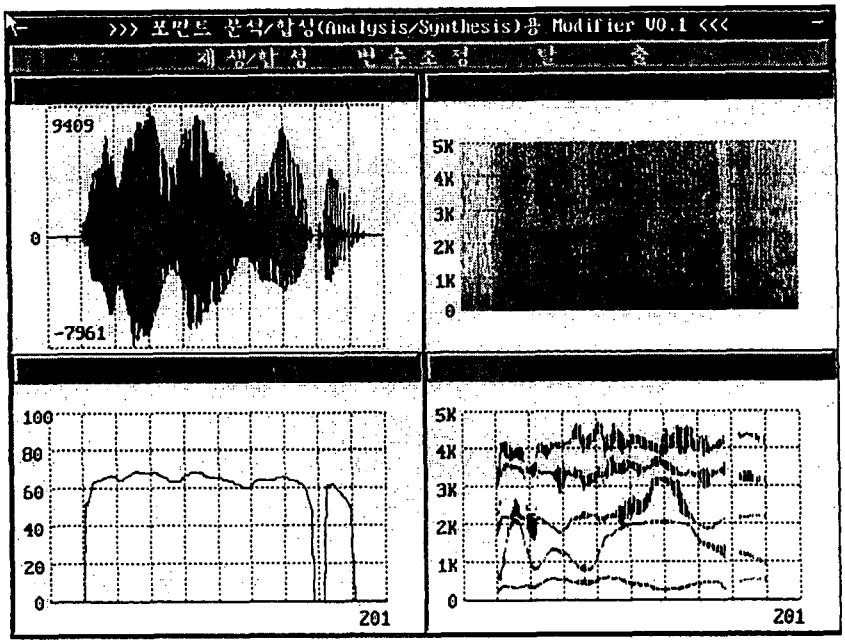


(c) Open quotien 70%

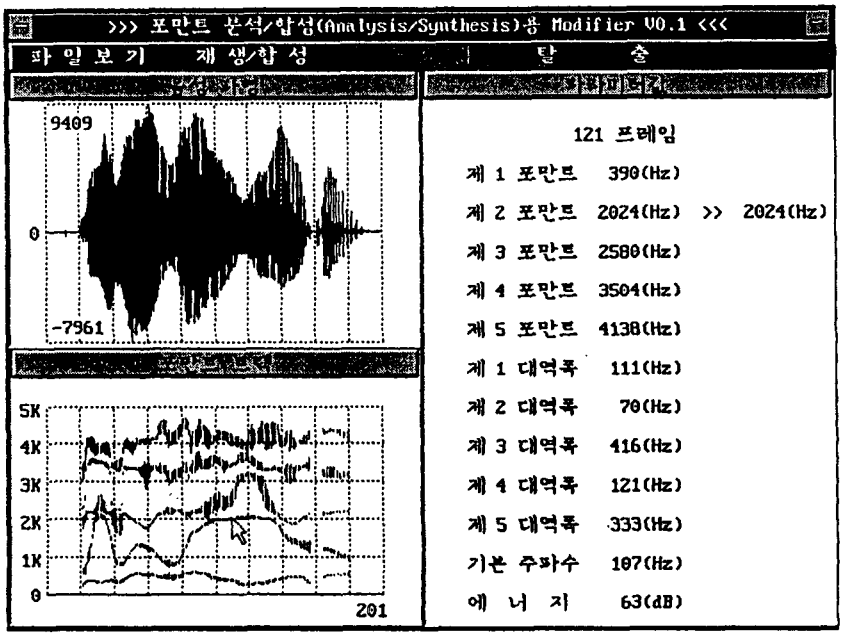


(d) Open quotient 90%

그림 10. 합성파라미터 표시 · 갱신 도구



(a) Display mode



(b) Modify · Synthesis moded

5. 결 론

이 연구에서는 음성신호의 포만트 분석 및 합성 시스템을 구현하였다. 정확한 포만트 정보의 추정을 위해 음성과 EGG 신호를 이용하여 유성음의 준주기적인 특성을 고려한 피치동기 방식과 closed phase 방식의 음성분석 방법을 이용하였는데 이러한 분석법에 필수적이면서도 음성만으로는 얻기 어려운 정확한 피치 정보를 EGG 신호를 이용함으로써 쉽게 구할 수 있었다. 그 결과, 음성과 EGG 신호를 이용한 2-채널 분석이 음성 외에 추가적인 신호가 수집·분석되어야 한다는 단점에도 불구하고 음성만을 이용한 1-채널 음성분석법의 성능평가 척도와 기준 모델을 제시할 수 있음을 확인할 수 있었다. 구해진 포만트 정보 및 피치 정보등의 합성파라미터들을 이용한 합성실험을 통해 음질, 파형 및 스펙트럼 특성이 원음에 가까운 합성음을 얻을 수 있었다. 또한, 포만트 합성방식에서 여기신호에 따른 합성음의 명료성은 LF 모델이, 자연성은 Rosenberg 모델이 뚜렷했었고, open quotient의 값이 클수록 부드러운 음질이 나타났으며, 값이 작을수록 선명한 음질을 보였다. 따라서 각 합성파라미터에 대한 음원의 모델링 및 파라미터 추출에도 이러한 요인들을 고려하여야 하며, 보다 나은 양질의 합성음을 위해서 이에 대한 연구가 앞으로 계속되어야 한다.

이 연구에서는, 특히, 합성파라미터의 변화가 합성음질 및 파형에 미치는 영향을 관찰하고자 합성파라미터와 음성파형 및 음성의 스펙트로그램을 화면에 나타내고 마우스를 이용하여 파라미터값을 사용자가 적절히 변경한 후 합성할 있는 합성도구를 개발하였다. 다양한 음성의 EGG 신호가 준비되어 분석·합성이 가능해 질 경우 개발된 도구를 이용한 파라미터 변화에 대한 연구는 포만트 합성 방식의 문자-음성 규칙합성기에서의 파라미터 제어방식의 개발과 적절한 합성단위 선정에 유용하게 사용될 수 있다.

참 고 문 헌

[1] K. S. Lee. 1992. Pitch synchronous analysis/synthesis using The WRLS-VFF-VT algorithm. Ph.D. dissertation. Univ. of Florida.

[2] R. L. Christensen., W. J. Strong, and E. P. Palmer. 1976. "A comparison of three methods of extracting resonance information from predictor-coefficient coded speech." IEEE Trans. ASSP, ASSP-24(1).

[3] S. S. McCandless. 1974. "An algorithm for automatic formant extraction using linear prediction spectra." IEEE Trans. ASSP, ASSP-22(2),

[4] B. Yegnanarayana. 1978. "Formant extraction from linear-prediction phase spectra." J. Acoust. Soc. Am. 63(5).

[5] 김웅식·배건성. 1990. "선형예측계수의 위상 스펙트럼의 미분치를 이용한 포만트 추정." 대한전자공학회 하계 종합 학술대회 논문집 13(1), 648~651.

[6] B. Yegnanarayana, D. K. Saikia, and T. R. Krishnan. 1984. "Significance of group delay function insignal reconstruction from spectral magnitude or phase." IEEE Trans.

ASSP, ASSP-32(3). [7] H. A. Murthy and B. Yegnanarayana 1991. "Formant extraction from group delay function," Speech Communication 10.

[8] D. H. Klatt and L. C. Klatt. 1990. "Analysis, synthesis, and perception of voice quality variations among female and male talkers." J. Acoust. Soc. Am. 87(2).

[9] E. L. Riegelsberger and A. K. Krishnamurthy. 1993. "Glottal source estimation : Method of applying the LF-model to inverse filtering." Proceeding of the International Conference on Acoustics, Speech and Signal Processing 2, 542~545.

[10] N. B. Pinto and D. G. Childers. 1989. "Formant speech synthesis : improving production quality." IEEE Trans. ASSP, ASSP-37(12).

[11] A. K. Krishnamurthy and D. G. Childers. 1986. "Two-channel speech analysis." IEEE Trans. ASSP, ASSP-34(4).

[12] 신무용 · 김정철 · 배건성. 1996. "음성 및 EGG 신호 분석에 의한 피치검출," 한국음향학회지 15(5), 5~12.

[13] J. L. Flanagan. 1972. "Voices of men and machines." J. Acoust. Soc. Am., 51, 1375-1387.

[14] G. Fant. 1959.. "The acoustics of speech." Proc. Third Internat. Congress on Acoustics, 188-201. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt. 1975. "Applications of a nonlinear smoothing algorithm to speech processing." IEEE Trans. ASSP, ASSP-23(6).

[15] J. N. Holmes. 1982. "Formant synthesizer: cascade or parallel?," JSRU Research Report 1017.

[16] J. N. Holmes. 1973. "The influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer." IEEE Trans. Audio Electroacoust, 198-305.

[17] D. H. Klatt. 1980. "Software for a cascade/parallel formant synthesizer." J. Acoust. Soc. Am. 67(3).

[18] 홍성훈 · 이정철 · 안수길 1992. "유성음의 성문파 추정." 제 9회 음성통신 및 신호처리워크샵 논문집, 59-73.

[19] 이준우 · 배건성. 1994. "음성 및 EGG 신호를 이용한 포만트 추정." 제7회 신호처리학회, 753~757.

본 논문은 한국과학재단의 핵심전문연구비(과제번호:941-0900-015-2) 지원으로 수행되었으며 지원에 감사드립니다.

접수일자 : '97. 1. 29.
게재결정 : '97. 2. 21.

▲ 이준우
서울시 서초구 우면동 16번지
LG 종합기술원 정보기술연구소 MI Gr.
Tel : (02) 526-7353(O) FAX : (02) 579-9781
e-mail : joonwoo@lgcit.com

▲ 손일권
경북 구미시 진평동 642 (우편번호 : 730-360)
LG 전자 TV O.B.U.
Tel : (0546) 470-2455(O) FAX : (0546) 470-2246
e-mail: ikson@lge.co.kr

▲ 배건성
대구광역시 북구 산격동 1370
경북대학교 전자전기공학부 (우편번호 : 702-701)
Tel : (053) 950-5527(O) FAX : (053) 950-5505
(053) 764-6991(H)
e-mail: ksbae@commlab.kyungpook.ac.kr