UIT Report Vol.13, No.1, pp.47~54, 1982 연구논문집 제 13권 제 1 호 pp.47~54, 1982

한국어 모음 음성의 합성에 관한 연구

이 태 호
전기 및 전자공학과
(1981.12.30 접수)

〈요 약〉

수개의 한국어 모음의 음성분석과 합성이 시도되었다. 분석에는 선형예측의 수단이 이용되었으며 그 결과를 수정하여 formant주파수와 대역폭을 얻었다. 이 정보들로부터 수 개의 2차계 여파기를 종속 연결한 발성기관의 모형이 구성되었으며 勵起信號로서는 충격파(impulse)를 저역 여파기를 통과시켜 사용하였다. 결과적으로 얻은 음성은 기계 음의 부자연스러운 특징이 뚜렷하였는데 勵起信號에 소량의 白雜音을 첨가함에 의하여 자연스러움이 크게 개선되는 것을 관측하였다.

A Study on the Systhesis of Korean Vowels

Lee, Taiho
Dept of Electrical and Electronic Eng.
(Received December 30, 1981)

〈Abstract〉

Some of the Korean vowels have been analyzed and synthesized. The LPC technique is used to analyze the natural vowels, and the result is modified to get formant parameters. The vocal tract is simulated by the cascaded second order filters representing formant resonators, and a lowpass filtered impulse is used for excitation of the vocal tract model. It is observed that the presence of a small amount of random noise in the excitation signal greatly improves the naturalness of the synthesized voice.

I. 서 론

음성의 합성 수단은 2개의 방향으로 크게 나눌 수 있는데 그 하나는 음성의 파형들을 저장하여 이를 적절히 연결하는 것이고 또 한가지는 발성기관의 모형을 세워 그 파라메타를 조직하는 방식이다.[1-4] 손쉽게 비교적 양질의 合成音을 얻는다는 점에서는 앞의 것이 유리하지만 기억장치의 규모나 傳送을 위한 회선의 경제성 그리고 융통성의 입장에서는 뒤의 방법이 절대적으로 우수하다. 영어를 비롯한 수개국어에 대해서는 두 방향 모두 상당한 연구가 이루어져 있어서 소박하나마 상품까지 나오고 있지만 한국어에 대해서는 아직 작업량이 매우 미미한 단계이다. 본 연구에서는 後者의 한 수단을 택하여 한국어의 모음 합성을 시도 하였다. 사용된 모형은 vocal tract의 공진주파수 즉, formant에 대응되는 공진여파기들을 종속연결한 것이며 이 여파기를 주기적으로 勵起시켜서 음성을 발생하는 것이다.

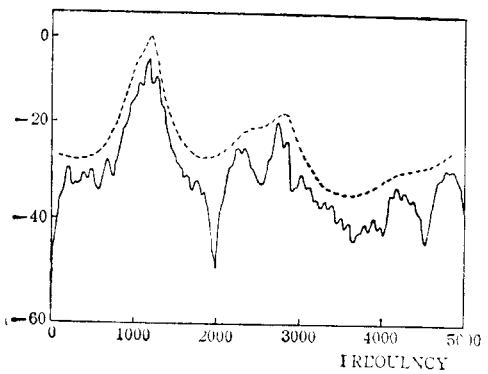
Formant에 관한 정보를 얻기 위해서 分析이 선행되었으며 여기에서는 선형 예측법이 사용되었다. 선형예측에 의한 분석의 결과는 formant의 대역폭을 정확히 결정해 주지 못하므로 직접 사용되지 않고 이 결과를 출발점으로 하여 합성—재분석—평가와 수정의 작업을 반복하는 지루한 과정을 거쳐서 최종의 formant를 결정하였다.

Ⅱ. 분 석

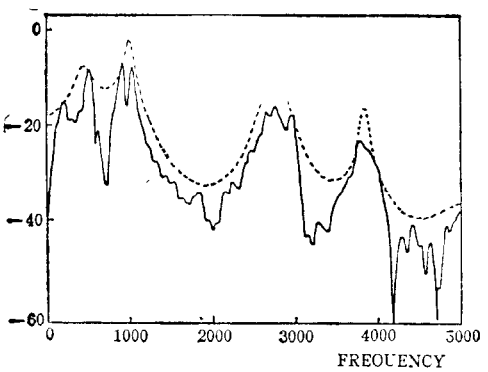
분석은 선형예측의 수단에 의하였는데 분석의 결과로 얻어지는 vocal tract의 전달함수는 다음과 같이 된다.[5]

$$H(z)=\frac{1}{1+\sum_{i=1}^{p} z^{-i}} \qquad (1)$$

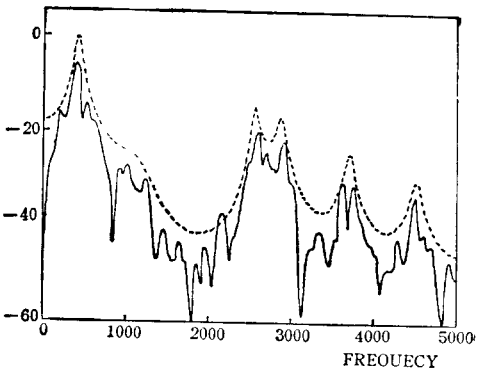
여기서 a_i들은 선형예측계수이며 p는 예측차수이다. 식 (1)은 극점들만으로 구성된 모형을 제공하여 한쌍의 공액극점들은 한 개의 formant의 후보가 된다. 만일 예측차수를 크게하면 실제 스펙트럼을 갈 재현해 주지만 잉여정보까지 포함하여 지나치게 많은 formant 후보를 제공하므로 바람직하지 못하다. 여기서는 $p=12$가 이용됐으며 계산은 autocorrelation 방법인 D'urbin의 공식이 사용되었다.



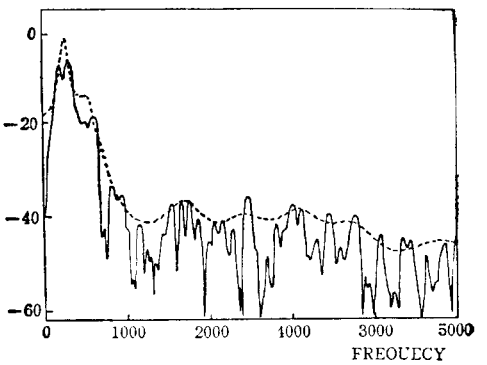
a. '아'



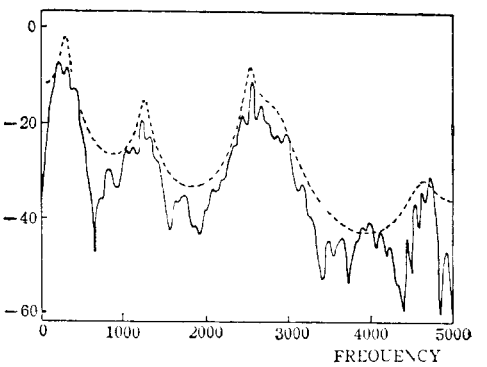
b. '어'



c. '오'

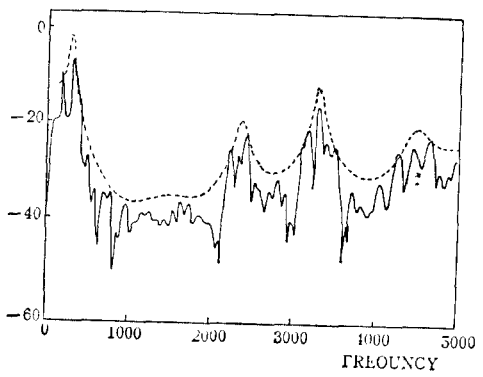


d. '우'

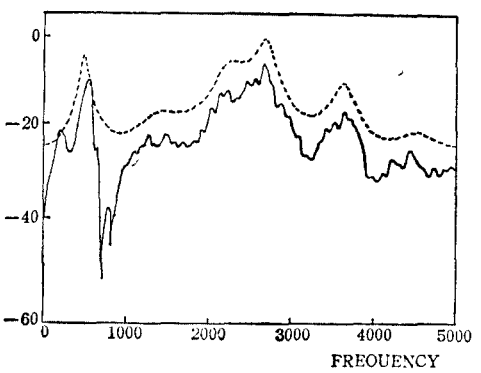


e. '으'

(그림 1. 다음면 계속)



f. '이'



g. '애'

그림 1. 자연음성의 스펙트럼

Bloken: 14th order LPC, Solid: DFT taken 5dB low, Vertical axis in dB scale.

분석에 앞서 전처리로서 먼저 음성신호를 저주파 여과기를 거쳐 3.4 kHz 이상을 잘라내고 10kHz의 율로 12bit로 샘플링하였다. 다음은 preemphasis를 거치게 하였는데 이것은$(1-z^{-1})$의 단순한 미분과정이며 preemphasis와 입술및 복사효과의 합성 효과는 2개의 영주파수 극점을 제거하게 되므로 결과적으로는 성대에서 발생하는 파형 즉, 여파기의 압력신호를 충격파로 가정할 수 있도록 해 준다.[6] 이것은 식 (1)이 vocal tract만의 모형으로 해석해도 된다는 것을 의미한다. 분석의 구간은 여성의 경우 3pitch, 남성의 경우 2pitch로 하여 대략 20 ms, 즉 200 샘플내외가 되도록 하였다. 각 구간은 Hamming window를 거쳐 분석에 들어가도록 하였다.

결과로 얻어진 여파기들의 주파수 특성을 自然音의 DFT(Digital Fourier Transform)와 함께 그림 1에 보였다. 그림에서는 $p=14$로 하였다.

Ⅲ. 합 성

1. Formant의 수정

선형예측방식에 의해서 발견된 formant 후보들은 최소한 다음 세가지 이유로 그 정확성이 제한된다. 즉, ① 주파수 특성은 window의 주파수 특성과 convolution이 이루어진 것이다. ② 수 개의 pitch를 한 분석구간으로 하므로 일종의 평균적 효과를 얻는다. ③ 한 pitch 內에서도 vocal tract의 특성은 변화를 보인다. 결과적으로 얻어진 스펙트럼은 주파수가 번지는 (spread)경향이 생기며 formant의 위치도 어느정도 부정확하다. 그렇다해도 잔류신호를 勵起用으로 사용하는 경우 상당히 좋은 품질을 얻을 수 있다. 그러나 잔류신호에 할당될 기억용량이 매우 커지게 되므로 모형을 세운 잇점이 소멸되어 vocoder가 아닌, 합성만을 전제로 하는 경우에는 단순한 형태의 일정한 勵起신호를 擇하는 것이 바람직하고 따라서 정선된 한 組의 formant를 필요로 하게 된다. 본 연구에서는 선형예측방법에 의하여 얻어진 formant 후보들을 수정함에 의하여 개선된 formant를 얻도록 하였는데 이 작업은 합성—결과의 분석—수정이란 다소 임의적이고 지루한 작업에 의하였으며 절차는 다음과 같았다. ① 자연음성의 DFT 및 선형예측분석. ② Formant후보로부터 합성음의 발생. ③ 합성음의 선형예측 분석. ④ 자연음의 DFT와 ③의 결과 비교. ⑤ 자연음과 합성음의 파형 비교. ⑥ 수정이 필요하면 ②로 되돌아 감.

이상의 과정에서 수정과 판단은 관찰에 의지하였으므로 평가의 척도가 없다. 따라서 다음의 두가지 기준을 만족시키는 범위에서 작은 수정만을 허용하였는데 그 하나는 formant의 대역폭을 가능한 한 줄여서 자연음의 파형에 근사한 감쇄특성을 얻도록 한다는 것과 또 하나는 스펙트럼의 형태를 유지하도록 노력한다는 것이었다. 위의 수정 작업은 한개의 프로그램에 격납되어 터미날에서의 대화에 의하여 연속적으로 반복작업이 이루어지도록 하였다. 결과는 표와 같다. (표의 기호는 다음절 참조)

표. 한국어 모음합성을 위한 formant

모 음	$F_1(\alpha_1)$	$F_2(\alpha_2)$	$F_3(\alpha_3)$	$F_4(\alpha_4)$	$F_5(\alpha_5)$	$F_6(\alpha_6)$
아	990(0.930)	1,215(0.975)	2,400(0.730)	2,630(0.968)	4,340(0.910)	
어	480(0.980)	990(0.973)	2,200(0.700)	2,770(0.956)	3,800(0.920)	4,300(0.500)
오	200(0.400)	415(0.986)	1,140(0.880)	2,560(0.960)	2,900(0.940)	3,690(0.970)
우	296(0.986)	520(0.807)	1,970(0.862)	2,760(0.850)	3,610(0.850)	4,626(0.900)
으	275(0.990)	1,230(0.985)	2,585(0.956)	2,820(0.770)	4,665(0.906)	
이	255(0.981)	2,340(0.960)	2,430(0.550)	3,300(0.960)	4,580(0.830)	
애	550(0.950)	2,043(0.870)	2,650(0.960)	3,610(0.880)	5,000(0.830)	

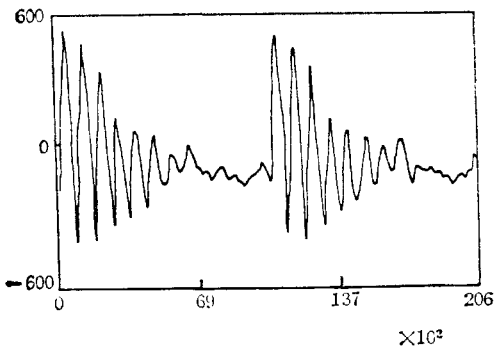
(F_k의 단위 : Hz)

2. 합성음의 발생

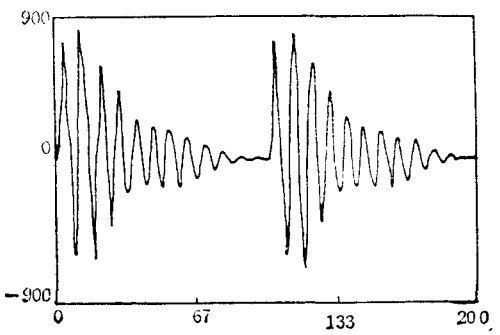
Formant가 결정되면 여파기를 구성하고 적절한 勵起신호를 인가하면 합성음을 얻을 수 있다. 여파기의 전달함수는 formant vocoder에서와 같이[7]

$$H(z)=\prod_{k=1}^{n}\frac{A_k}{1-2e^{-\alpha_k T}\cos(2\pi F_k T)z^{-1}+e^{-2\alpha_k T}z^{-2}} \qquad (2)$$

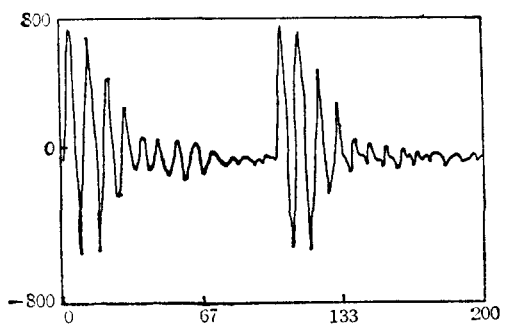
로 되며 여기서 F_k와 α_k는 각각 k번째 formant의 주파수와 감쇄정수이며 T는 샘플링간격(10^{-4}sec), A_k는 k번째 공진기의 이득이 해당 formant주파수에서 단위의 값이 되도록 조절한 이득이다. 보통 formant vocoder에서는 어떤 모음을 특징지우는데 3개의 formant가 필요한것으로 알려져 있으며 여기에 1개의 고정된 formant를 높은 주파수 영역에 추가하여 전체적인 스펙트럼의 형태에 균형을 취하는 수단이 쓰인다.[7] 그러나 여기서는 5개 또는 6개의 formant를 사용하였는데 그 까닭은 아직 한국어 음성의 통계적 성질이 파악되기 못했기 때문이었다.



a.



b.



c.

그림 2. 모형 '아'의 파형

a. 자연음성,
b. 합성음성—pitch의 끝부분에 추가적 감쇄를 준 것,
c. 합성음성—백잡음 입력을 추가한 것

勵起신호는 전처리 과정의 preemphasis를 보상하기 위하여 $(1-0.95z^{-1})^{-1}$에 의하여 deemphasis를 거친 충격파를 사용하였는데 $(1-z^{-1})^{-1}$을 사용하지 않은 것은 직류분의 축적을 억제하기 위해서였다. 결과로 얻은 파형은 자연음과 비교하였을 때 끝 부분에서 차이를 보인다. 자연음의 경우는 한 pitch의 끝부분에서 진폭이 급격히 감쇄하는 특성을 보이기 때문이다. 이것은 성대가 열리는 기간 발성기관이 폐와 연결되는 효과 때문인 것으로 해석하여 한 pitch의 뒤쪽 35%기간에 추가적 감쇄를 시도한 결과 자연음에 근사한 파형을 얻었으나 (그림 2의 b) 청취시험 결과는 매우 실망스러운 것이었다. 약간 鼻音化되는 것이 그 특징이었는데 鼻音의 경우 鼻腔에 의한 감쇄가 일어나는 것과 유사한 효과가 일어나고 있는 것으로 해석된다.

추가적인 감쇄를 가장 지배적인 formant에만 적용했을때는 음질이 별로 변화를 보이지 않았는데 어느 경우에나 기계음의 부자연스러움을 없앨 수는 없었으나 각 모음의 특징은 뚜렷이 확인할 수 있었다. 충격파 에너지의 2.5%의 백잡음을 勵起신호에 추가함에 의하여 自然스러움은 크게 개선할 수 있었는데 이 때의 파형은 그림 2의 c에서와 같이 한 pitch의 뒷부분에서 자연음과 매우 유사하게 되었음을 알 수 있었다.

주파수 특성은 그림 3과 같으며 다른 모음의 경우는 그림 4에 자연음과 함께 보여졌다.

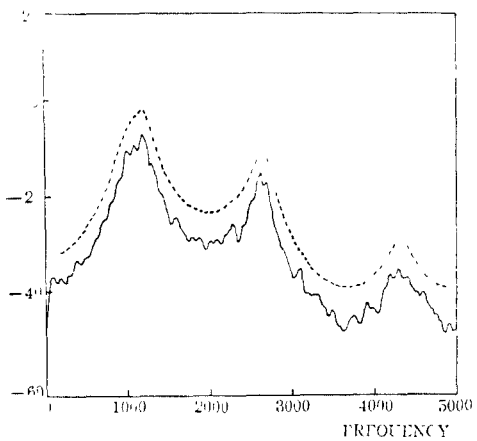
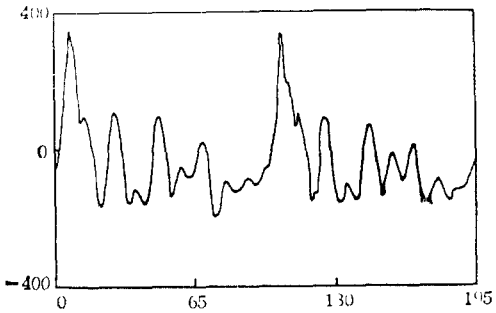
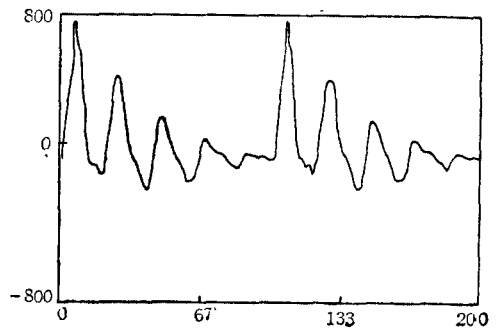


그림 3. 합성음 '아'의 스펙트럼

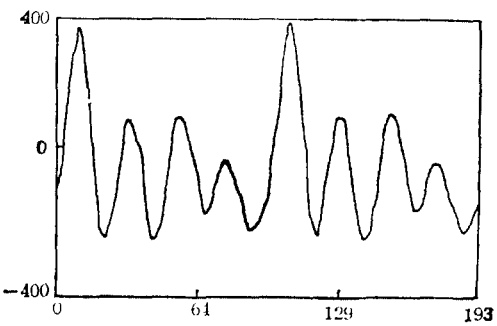
Broken: 14th order LPC, Solid: DFT taken 5dB low, Vertical axis in dB scale.



a. 자연 '어'

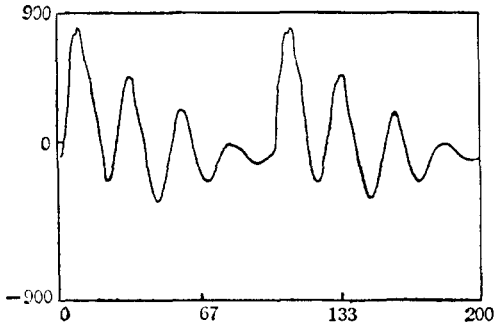


b. 합성 '어'

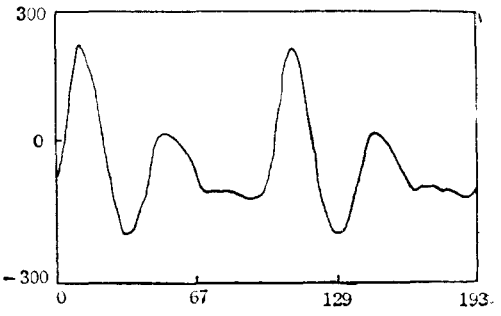


c. 자연 '오'

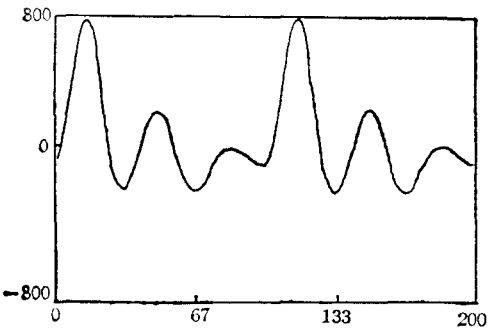
(그림 4. 다음 면에 계속)



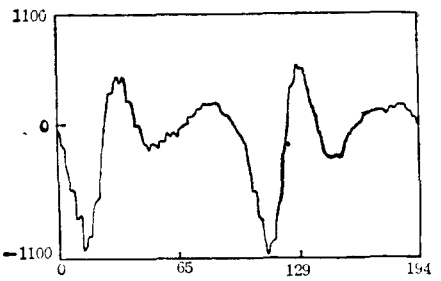
d. 합성 '오'



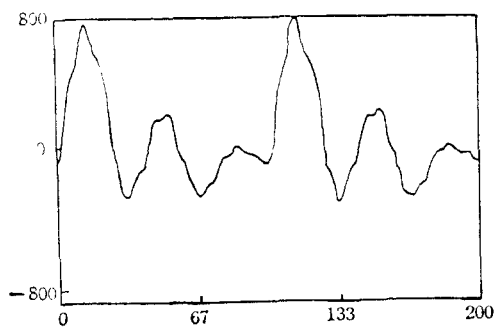
e. 자연 '우'



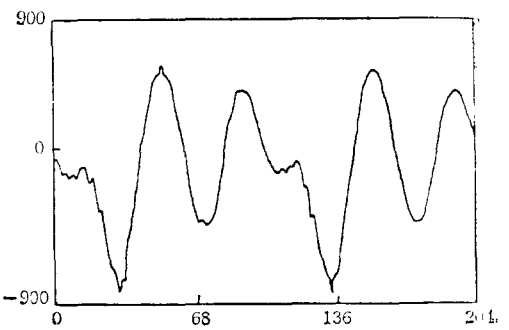
f. 합성 '우'



g. 자연 '으'

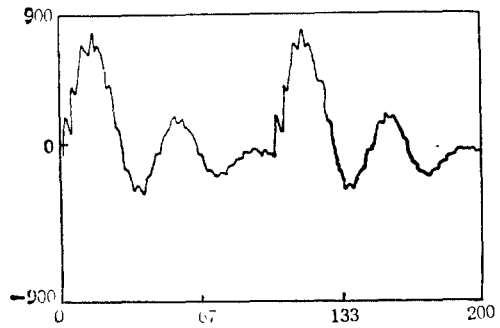


h. 합성 '으'

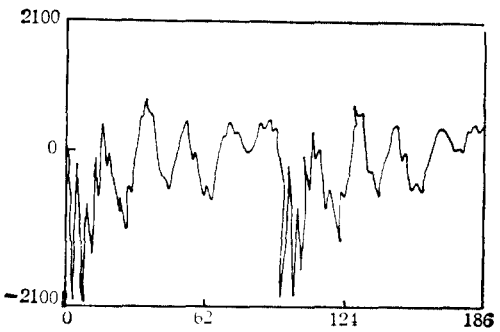


i. 자연 '이'

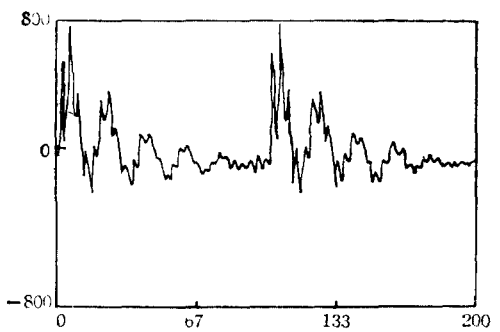
(그림 4. 다음 면에 계속)



j. 합성 '이'



k. 자연 '애'



l. 합성 '애'

그림 4. 자연음성과 합성음성의 파형 비교
(수평축-sample 수, 수직축은 임의로 잡음)

Ⅳ. 결 론

어느정도 만족할만한 합성음을 얻기는 하였으나 여러 단계에서 시행착오를 채용하고 있어서 잘 발달된 논리적 배경에 토대를 두고 있음에도 본 연구의 합성과정이 하나의 숙련기술의 한계에서 크게 벗어나지 못하고 있다는 인상을 준다. 몇가지 제기될 만한 문제들을 토의해 보기로 한다. 첫째로 formant를 수정하는 과정은 정량적 척도가 없다. 이 문제의 상당량은 통계적 자료가 축적됨에 따라 경감될 수 있다. 그러나 음성의 최종적 평가는 청각에 의한 시험을 거쳐야 하는 것이므로 이 단계의 완전한 배제는 어려운 것이다. 또하나의 관심의 대상은 추가적 감쇄의 영향이다. (그림 2의 b) 매우 작은 에너지의 변화에 의한 청각에의 효과가 이렇게 클 수 있다는 것은 놀라운 일이다. 이것은 파형의 낮은 에너지 부분 즉 한 pitch의 끝 부분이 실제로 청각에 미치는 영향이 적지않음을 암시한다. 이것은 다른 문제, 즉 잡음 혼입에 의한 효과에서도 재확인 되며 미량의 잡음이 음성 품질의 개선에 상당한 영향을 준다는 것은 勵起신호에 관한 상세한 평가를 요청한다. Glottal wave의 형태는 $\sin^2(at) \rightarrow \cos(b+ct)$의 형태 (a, b 및 c는 적절한 상수)가 정설로 취급되어 오고 있으나[8] 이 모형에 의한 합성의 결과는 여기서 사용한 2중적분된 충격 파형에 비하여 개선의 점이 발견되지 않았다. Matausek 등에 의한 영점과 극점의 복합 모형은 흥미있는 시도이지만[6] 모형의 복잡성에도 불구하고 뚜렷한 개선이 기대되지 않는다. 그 까닭은 이들이 모두 본 연구의 백잡음과 같은 고역특성을 갖지 못할 뿐 아니라 자연음 또는 백잡음 입력의 경우에서 볼 수 있는 바와 같이 낮은 에너지 부분에서 갑자기 나타나는 작은 고역 신호들을 발생시킬 수 없기 때문이다.

Wong 등의 수단[9]에 의하여 잔류신호를 하나 또는 수개의 표준형으로 압축한다면 개선된 勵起신호로 쓸 수 있을 것이다.

참 고 문 헌

1. L.H. Rosenthal, et al, "A multiline computer voice response system utilizing ADPCM coded speech", IEEE Trans. on Acoustics, Speech, and Signal Proc., vol. ASSP-22, pp. 339—352, Oct. 1974.
2. L.R. Rabiner and R.W. Schafer, "Digital techniques for computer voice respoese: Implementations and applications", Proc. IEEE, vol. 64, pp. 416—433, Apr. 1976.

3. J. Allen, "Synthesis of speech from unrestricted text", Proc. IEEE, vol. 64, pp. 433—442, Apr. 1976.
4. J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Synthesis of speech from a dynamic model of the vocal chords and vocal tract", BSTJ, vol. 54, pp. 484—506, 1975.
5. J. Makhoul, "Linear prediction: A tutorial review", Proc. IEEE, vol. 63, pp. 561—580, Apr. 1975.
6. M. R. Matausek and V. S. Batalov, "A new approach to the determination of the glottal waveform", IEEE Trans. On Acoustics, Speech, and Signal Proc., vol. ASSP-28, pp. 616—622, Dec. 1980.
7. L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, pp. 379, Prentice Hall, New Jersey, 1978.
8. H. W. Strube, "Determination of the instant of glottal closure from the speech wave", JASA, vol. 56, pp. 1625—1629, 1974.
9. D. Y. Wong, J. D. Makhoul, and A. H. Gray, JR., "Least squares glottal inverse filtering from the acoustic waveform", IEEE Trans. on Acoustics, Speech, and Signal Proc., vol. ASSP-27, pp. 350—355, Aug. 1979.