한국어 문어변환 시스템 내에서의 음성합성기 개발

The Development of Speech Synthesizer In Korean TTS System

姜 贊 熙*, 陳 庸 玉**

(Chan Hee Kang*, Yong Ohk Chin**)

요 약

본 논문은 매 40ms 정도의 음성파형으로 부터 추출된 6 내지 9ms 정도의 1 피치주기 파형을 합성단위로 사용하여 합성시킨 시간영역에서의 합성방식을 한국어 문어 변환 시스템 내에서의 음성합성기에 적용시킨 연구결과이다. 실험결과, 4가지 유형의 한국어 음절 합성이 가능하고, 장단강약과 같은 운율요소의 제어가 용이하고, 또한 합성 알고리즘이 간단하여 실시간 처리가 가능하였으나, 문장 단위의 음성을 합성하기 위하여는 문장내에서의 다양한 피치 패턴에 대한 연구와 이의 효율적인 제어에 관한 연구가 이루어져야 할 것이다. 합성음에 대한 평가방법으로는 원음과 합성음에 대한 시간영역에서의 파형비교, 주파수 영역에서의 스펙트럼 포락선 유사성 비교 및 합성음에 대한 청취도 실험을 행하였다.

ABSTRACT

This paper is the results applied the synthesis method in time domains with 1 pitch period waveforms of 6 or 9ms in lengths extracted from short time intervals(about 40ms in lengths) to the speech synthesizer in korean TTS system. By the results of experiment, it is possible to synthesize 4-type korean syllables, easy to control the parameters for prosodic elements such as durations and stresses and also possible to process with real time because the syntesis algorithm is too simple. But to synthesize the sentences we have to investigate and establish various pitch patterns in a sentence and to develop the synthesis algorithm which is possible to efficient control of them. And we had the validity test through the waveform comparision with original speeches and synthesized speeches, the resemblances of spectrum envelop and hearing test to them.

I. 서 론

한국어 문어전환(TTS : Text-to-Speech) 시스템을 개발하기 위하여 선결하여야 할 과제 중의 하나가 문장 단위의 무제한 합성이 가능한 음성합성기(speech synthesizer)의 개발[1]로써, 본 논문에서는 CYBEX 시스템[2](그림 1)내 1 피치 프레임 단위의 음성파형에 의한 무제한 합성방식을 이용한 TTS 음성합성상치를 개발하기 위한 제 1단계의 연구과정으로 한국어 음절유형별로 합성시킨 결과에 대하여 주로 다루었으며, 이 방식을 제한된 문장에 적용시킴으로써 합성방식에 대한 타당성과 TTS 시스템내 음성합성기로 사용되기 위한 보완점 및 앞으로의 연구방향 설정에

* 상지대학교 병설 전문대학 전자과
** 경희대학교 전자공학과
접수일자 : 1992년 11월 27일

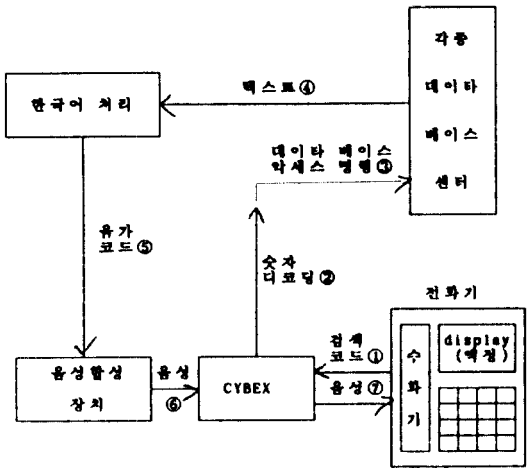


그림 1. CYBEX 시스템내에서의 한국어 문어변환 장치 블록도

Fig 1. The block diagram of korean TTS in CYBEX system

관하여 논하였다. 일반적으로 문장단위의 무제한 합성을 위하여는 음질이 양호하여야 할 뿐만아니라 자연스런 합성음 즉, 주어진 문맥에 따라 운율요소(prosodic elements)의 제어가 자유로운 합성 방식이어야 한다. 지금까지 연구 발표되어진 국내외의 TTS시스템에서의 음성합성 방식은 주파수영역에서의 합성방식이 주류를 이루고 있는데[3][4] 이는 시간영역에서의 음성합성 방식에서는 합성에 이용되는 합성단위에 따라 음질과 운율요소의 제어가 용이한 음성합성기를 개발하기에는 아직도 극복하기 어려운 제약점이 많아 실용화가 어려운 반면에, 주파수 영역에서의 합성방식은 시간영역에서의 음성합성 방법보다 음질이 저하되기는 하지만 운율요소의 제어가 가능한 합성방식이 개발되어 문장단위의 음성합성이 가능하기 때문이다. 일반적으로 TTS 시스템내 음성합성기로 사용되는 주파수 영역에서의 음성합성방식은 자연언어(natural language) 처리를 행한 후, 그때의 구문 성분에 해당하는 피치 패턴을 데이타 베이스화시켜 변환된 구문의 음소열에 해당되는 LSP (Line Spectrum Pairs)에 피치 패턴을 인가하여 문장단위의 음성을 합성시키는 LSP를 이용한 방식[3]과 자연언어 처리에 의한 음소열을 발생시켜 인간의 조음기관과 가장 유사한 직병렬 포르만트 합성기로 합성음을 발생시키는 MITalk 시스템[1]등을 대표적인 예로 들 수 있다. 그러나 시간영역에서의 음성합성기는 주파수 영역에서의 합성방식에 비하여 합성방식이 간단하고 음질이 양호하다는 장점을 지니고 있는 반면에 자연스런 합성음을 발생시키기 위하여는 합성에 이용되는 단위길이가 길수록 좋으나 데이타 베이스양이 기하급수적으로 증대하는 폐단이 있다. 즉, 문장을 합성 단위로 사용하면 음질과 자연성은 뛰어나지만 임의로 문장내의 운율이 변하는 다양한 문장의 무제한 합성은 불가능하여 극히 제한된 문장단위의 합성만 가능하다. 또한, 단어(words)나 어절(phrases) 혹은 음절(syllables) 단위로 저장시켜 합성할 경우에는 음절 단위의 합성음은 음질이 뛰어나기는 하지만 저장된 형태의 음성을 편집, 재생시키므로 상단, 고저, 강약, 억양등과 같은 다양한 변화가 있는 자연스런 문장 단위의 합성은 기대하기 어려우며, 음소와 음소, 음절과 음절구간이 만나는 천이영역에서의 조음변화 처리가 어려워 제한된 경우에만 실용화되고 있다. 음절단위 보다 더 적은 음소(phoneme) 단위의 파형을 메모리에 저장하여 합성시킬 경우에는 무제한 합성이 가능하나, 음절 이상의 합성단위를 사용하였을 때 보다 규칙합성(synthesis by rule)에 사용되는 매개변수의 사용량이 급증할 뿐만아니라 합성음질이 나빠서 실시스템에서는 거의 사용되고 있지 않다[5]. 일례로, 포래스트 모저(Mozer)[6]가 제안한 방식에서와 같이 음소단위에 의한 합성방식에서는 영문에서 "s", "k", "u:", "l" 등과 같은 50 내지 200ms 정도의 자음소와 모음소에 해당되는 음소 200개 정도를 저장시킨 후 이들 파형을 연결시켜 "school"이란 음성을 합성시키므로, 이는 저장된 음소파형을 녹음 재생시킨 결과가 되어 문장단위의 합성시 운율 요소의 다양한 변화가 어려울 뿐만아니라, 음소와 음소구간이 만나는 천이영역에서의 불일치 등으로 음질 또한 나빠 문어전환 시스템에서 사용하기에는 부적합하다. 따라서 본 논문에서는 이와같은 시간영역에서의 음성합성방식이 지니고 있는 한계점을 극복하기 위하여 합성단위로 매 40ms 정도의 음성파형으로 부터 추출된 6내지 9ms 정도의 1 피치주기 파형을 선정하여 합성시킨 과정 및 결과를 컴퓨터 시뮬레이션으로 제시하였다. 본 논문에서 행해진 과정을 개략적으로 설명하면, 2장에서는 한국어 문어변환 장치의 시스템 개요 및 합성에 이용되는 10ms 단위의 무성자음 데이타 화일과 규칙합성용 데이타 포맷 화일에 대하여 논하였으며, 3.4절에서는 규칙합성용 매개변수 추출과정과 이를 이용하여 합성시킬 경우 발생되는 스파이크성 잡음 제거에 대하여 설명하였다. 3장에서는 단음절 단위를 여러가지 형태로

규칙합성시킨 예와 한국어 단모음 및 이중모음에 적용시킨 합성음의 결과를 제시하였으며, 또한 시간영역에서의 합성방식을 제한된 문장에 적용시킨 결과를 컴퓨터 시뮬레이션으로 제시하여 이 방식에 대한 타당성을 검토하였다.

Ⅱ. 한국어 문어 변환 장치

1. 시스템 개요

본 논문에서의 한국어 문어 변환 장치란 입력된 텍스트 정보로 부터 자연언어 처리를 거쳐 생성된 음소 변환 기호열에 의거 합성음을 발생시킴을 뜻하는 전용회선(private line)이나 공중선 또는 공중전화 등을 통하여 데이타의 전송과 수신이 이루어지는 장치를 말하며, 시스템 구성은 그림 2에서와 같이 자연언어처리부와 음성합성처리부 및 데이타 송수신부로 구성된다. 자연언어 처리부에서는 입력되어진 임의의 문자열을 2 바이트 조합형 코드로 변환시킨 후 음운변동 규칙 및 운율생성 규칙에 따라 악센트 유형, 지속시간, 피치, 진폭 및 자음과 모음의 세그먼트 변지와 각 음절의 운율과 같은 매개변수들을 결정하여 새로운 부호열의 음가 코드를 생성시켜 음성합성 처리부로 보내면 부성자음 데이타 화일과 규칙합성용 1,096 음절 데이타 포멧 화일에 저장되어 있는 규칙합성용 매개변수를 이용하여 프레임 단위로 분석화시켜 합성음을 발생시킨다.

2. 무성자음 데이타 화일과 규칙합성용 데이타 포멧 화일

표 1에서와 같이 한국어의 자모음은 1989년 3월 시행된 표준어 규정에 의하면 표준어의 자음은 19개(ㄱ,ㄲ,ㄴ,ㄷ,ㄸ,ㄹ,ㅁ,ㅂ,ㅃ,ㅅ,ㅆ,ㅇ,ㅈ,ㅉ,ㅊ,ㅋ,ㅌ,ㅍ,ㅎ)로, 표준어의 모음은 21개(ㅏ,ㅐ,ㅑ,ㅒ,ㅓ,ㅔ,ㅕ,ㅖ,ㅗ,ㅘ,ㅙ,ㅚ,ㅛ,ㅜ,ㅝ,ㅞ,ㅟ,ㅠ,ㅡ,ㅢ,ㅣ)로 규정하고 있다. 이 중에서 초성의 위치에는 자음 19개가, 중성의 위치에는 단모음 10개와 이중모음 11개가 올 수 있으며, 종성의 위치에는 자음 7개가 올 수 있다. 이들이 결합하여 1개의 음절을 구성하며, 이때 음절 수는 V형이 21개, VC형이 147개, CV형이 399개, CVC형이 2,793개로 모두 3,360개가 존재하나, 실제 주로 사용되는 음절 수는 1,096음절 정도이다. 무성음인 경우에는 10khz 샘플링 비로 획득한 음성 데이타의 파형을 분석하여 10ms 단위의 파형을 저장하여 무성 자음에 대한 데이타 세그먼트 화일을 생성하였으며, 유성음인 경우 즉, 피치의 준주기성이 나타나는 파형의 구간에서는 음절당 매 40ms 간격으로 추

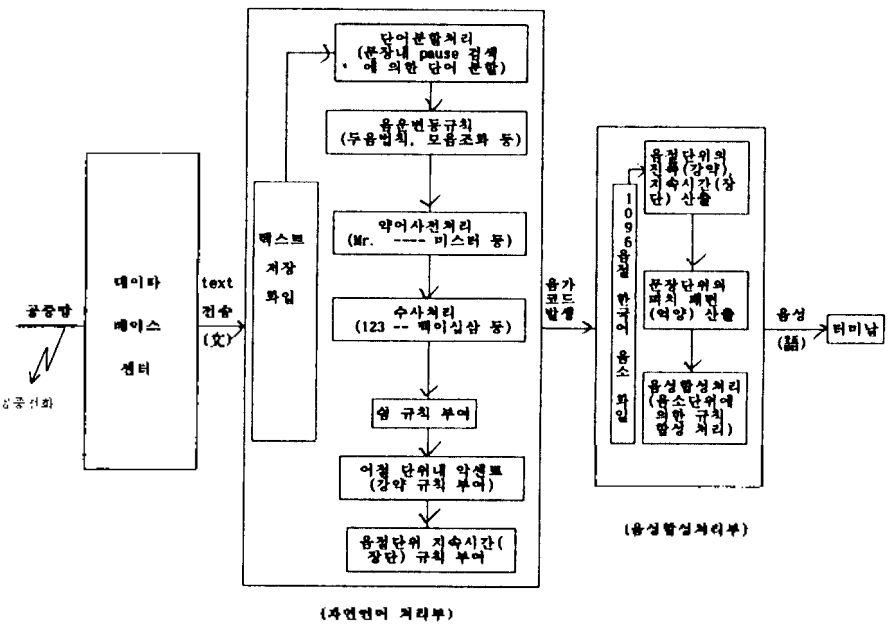


그림 2. 한국어 문어 변환 장치 구성도
Fig 2. The plot of Korean TTS system configuration

표 2-1. 한국어의 자모음 분류
Table 2-1. The classification of Korean consonants and vowels

			입술소리	혀끝소리	구 개 음	연구개음	목청소리
자음(19개)	파열음	예사소리	ㅂ	ㄷ		ㄱ	
		된 소 리	ㅁ	ㄸ		ㄲ	
		거센소리	ㅍ	ㅌ		ㅋ	
	파찰음	예사소리			ㅈ		
		된 소 리			ㅉ		
		거센소리			ㅊ		
	마찰음	예사소리		ㅅ			ㅎ
		된 소 리		ㅆ			
		거센소리					
	비음		ㅁ	ㄴ		ㅇ	
	유음			ㄹ			
모음(21개)	단 모 음		ㅏ, ㅓ, ㅗ, ㅜ, ㅡ, ㅣ, ㅐ, ㅔ, ㅚ, ㅟ				
	이중모음		ㅑ, ㅕ, ㅛ, ㅠ, ㅒ, ㅖ, ㅙ, ㅞ, ㅢ, ㅘ				

출한 1 피치 프레임 구간의 음성데이타와 진폭, 지속시간, 피치주기 등의 매개변수를 추출하여 1,096 음절에 대한 규칙 합성용 데이타 포맷 화일을 생성하였다.

3. 규칙합성용 매개변수 추출

음소단위에 의한 규칙합성시 텍스트로 부터 변환된 음소 기호열에 따라 저장된 음성 데이타를 악세스하여 합성시킬 경우에는 운율요소의 제어가 용이하지 못하여 서론에서와 같이 자연스런 합성음을 발생시킬 수 없다. 따라서 본 논문에서는 이와같은 문제점을 해결하기 위하여 규칙합성시 파라메타의 제어가 용이하도록 1 피치 구간의 음성파형을 추출하여

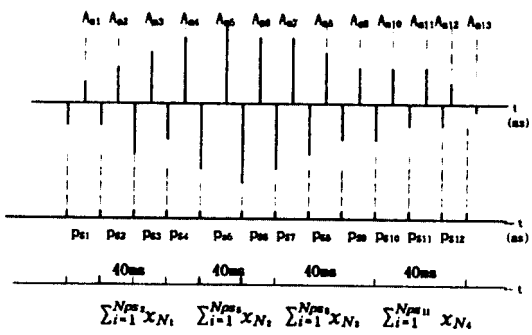


그림 3. 규칙합성용 매개변수
Fig 3. Parameters for the synthesis by rule

이를 합성단위로 사용하여 합성시켜 운율요소의 제어를 시도하였다. 이를 위하여 합성시키고자 하는 음절의 파형을 분석하여 진폭, 지속시간과 피치주기등과 같은 성분을 합성용 매개변수로 추출시켜 규칙합성하였으며, 이를 이용한 결과는 3장에 제시하였다. 음성파형으로 부터 규칙합성용 매개변수를 추출하기 위하여 그림 3에서와 같이 임의의 음성 데이타를 x(n), 단음절의 데이타 갯수를 N, 단음절내에서의 1 피치주기의 프레임 갯수를 N_P로 각각 정의하면 단음절의 음성 데이타 열은 $\sum_{n=1}^{N} x(n)$으로 표기된다. 이때 각각의 피치 프레임 구간의 경계를 그림 3에서와 같이 P_{s1}, P_{s2}, P_{s3}, … 등으로 나타내고, 각 피치 프레임 구간에서의 데이타 갯수를 N_{Ps1}, N_{Ps2}, N_{Ps3}, … 등을 배열 $N_{Ps}(\ )$로 표기하면, N 개의 음성 데이타 열 $\sum_{n=1}^{N} x(n)$을 1차원 배열인 1 피치 프레임 단위의 N_P 개 소블록의 합으로 표기 가능하므로 이를 2차원 배열로 표시하면,

$$\text{즉, } \sum_{i=1}^{N} x(n) = \sum_{n1=1}^{Np} \sum_{n2=1}^{Nps(n)} x(n1, n2) \qquad (1)$$

(단, $n = n1 + \sum_{n1}(n1-1)\cdot N_{ps}(n1-1)$, $N_{ps(0)}=0$ 임.)

와 같다. 여기서, x(5, 10)은 5번째 피치 프레임 열의 10번째 데이타를 의미한다. 또한, 단위 피치 프레임 구간내에서의 음성 데이타 열의 최대 진폭의 절대치를 각각 A_{m1}, A_{m2}, A_{m3}, … 등으로 정의하고, 각각의 피치 프레임 단위 구간내의 데이타 열을 일정한 크기

로 정규화시킨 임의의 음성 데이타를 $x_N(n)$로 정의하면, 2차원 블록화 배열로 표시된 음성 데이타 $\sum_{n1=1}^{Np}\sum_{n2=1}^{Nps(i)} x(n1,n2)$은

$$\sum_{n1=1}^{Np}\sum_{n2=1}^{Nps(i)} x(n1,n2) \approx \sum_{n1=1}^{Np}\sum_{n2=1}^{Nps(n1)} A_m(n1)\cdot x_N(n1,n2) \quad (2)$$

로 표시된다. 또한 단음절내 40ms 구간에서의 피치 갯수를 $M_1, M_2, M_3, \cdots, M_{N_B}$(여기서, 10khz의 샘플링 비로 데이타를 획득한 경우에는 $N_B = INT(N/400)$임)로 정의하고 이를 1 차원 배열 M( )로 표시하면,

$$N_P \approx M(1)+M(2)+\cdots+M(N_B) \quad (3)$$

이 되며, 1 피치 프레임 단위의 합으로 표시된 $\sum_{n1=1}^{Np}\sum_{n2=1}^{Nps(n1)} A_m(n1)\cdot x_N(n1,n2)$를 식(2.3)에 의하여 40 ms 단위로 구성된 N_B개의 2차원 배열의 블록 합으로 표기 가능하므로 이를 3차원 배열의 식으로 표시하면

$$\sum_{n1=1}^{Np}\sum_{n2=1}^{Nps(n1)} A_m(n1)\cdot x_N(n1,n2) \approx \sum_{n1=1}^{N_B}\sum_{n2=1}^{M(n1)}\sum_{n3=1(단, M(0)=0임)}^{Nps(2+\sum(n1-1)\cdot M(n1-1))} A_m(n2+\sum_{n1}(n1-1)\cdot M(n1-1))\, x_N(n1,n2,n3) \quad (4)$$

이 되며, 이 식에서 $x_N(n1,n2,n3)$는 x_N(단음절내 블록번호, 블록내 단위피치 프레임 번호, 단위피치 프레임내 데이타 번호)를 의미한다. 윗 식들로 부터 매개변수를 사용한 규칙합성용 매개변수들을 정리하여 보면, 식(1)에서 추출된 매개변수는 N_P+1개(이는 이타 갯수 정보 $\sum_{i=1}^{Np} N_{Ps}(i)$, $(=N_{Ps}(1), N_{Ps}(2), \cdots, N_{Ps}(N_P))$로써 N_P개임)와 식(2)에서 추출된 단위피치 구간내에서의 최대진폭 정보 N_P개($\sum_{i=1}^{Np} A_m(i)$)와 식(3)에서 추출된 N_B개의 40ms 단위 블록 내에서의 단위피치 프레임 개수 정보($\sum_{i=1}^{N_B} M(i)$)와 식(4)에서와 같이 40ms 단위 블록 내에서 추출된 2번째의 단위 피치 프레임 정보($\sum_{i=1}^{Nps(N_B)} x_N(i)$)등 이며, 이 과정을 흐름도로 표시한 것이 그림 4이다.

임의의 특정화자로 부터 입력된 단음절중에서 정상상태의 모음발성구간의 정상부분을 찾아 단음절내에서의 피치주기를 탐색한 후 파형의 엔벨로프(진폭), 지속시간과 유무성음식별과 같이 분석된 성분을 저장시켜 규칙합성에 이용하였으며, 분석된 파형의

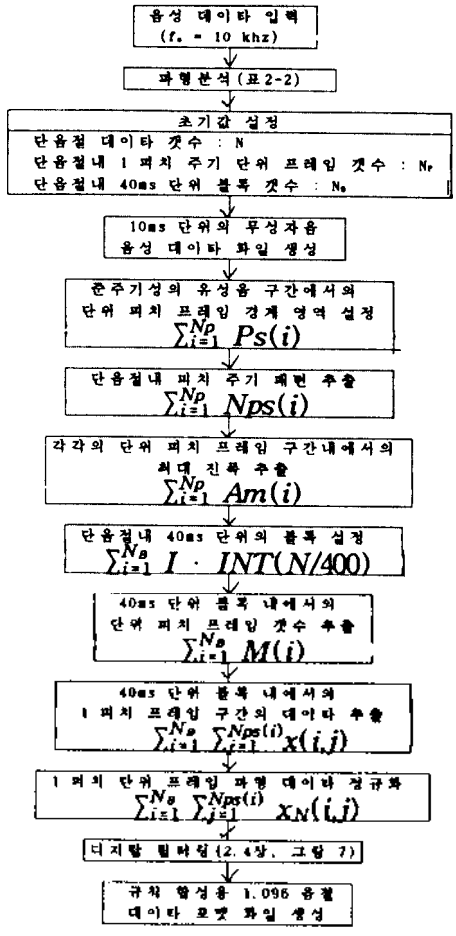
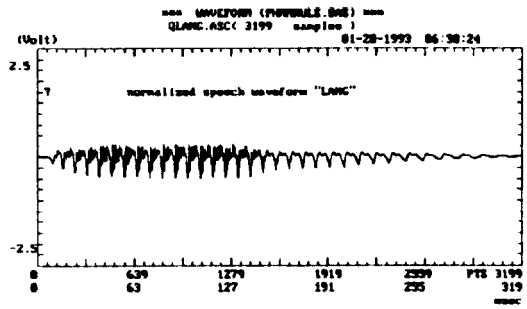


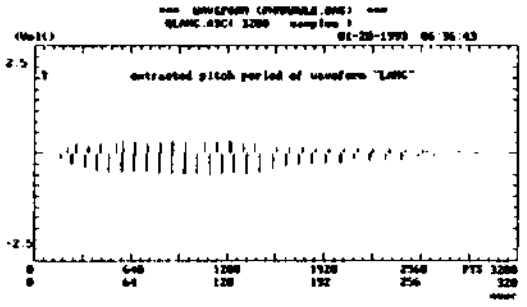
그림 4. 규칙합성용 매개변수 추출 흐름도
Fig 4. Extraction flowchart of parameters for synthesis by rule

결과를 표2-2에 제시하였다. 표2-3은 표2-2로 부터 구한 피치주기를 표시한 것이다. 그림 5(b)는 시간영역에서 합성시키고자 하는 기본 파형의 입력된 사전정보이며, 이를 이용하여 규칙합성시킬 경우 합성 규칙에 따라 여러가지 패턴의 파형으로 변화시킬 수 있으며, 합성 결과는 3장의 그림 9에 제시하였다. 문장단위의 합성을 위하여는 단음절에 대한 파형 분석 보다는 2 음절어 이상의 문장성분에 따른 진폭 패턴, 지속시간 패턴, 피치 패턴, 악센트 패턴과 쉬(pause) 패턴등을 분석하여 합성에 이용하여야 음절단위로 합성시켰을 때 발생되는 부자연스런 현상을 제거시킬 수 있을 것이다. 이 방식에 의한 문장단위의 합성 가능성을 타진하기 위하여 극히 제한된 문장("가는 말이고와야 오는 말이 곱다.")을 분석하여 합성시킨 결

파를 그림 12와 그림 13에 제시하였다.



(a)



(b)

그림 5. (a)단음절 "랑"의 시간파형
(b)규칙합성을 위한 기본파형
Fig 5. (a)Speech waveform "lang"
(b)Basic waveform for the synthesis by rule

표 2-3 추정된 피치주기
Table 2-2. Estimated pitch period

SPEECH SIGNAL (LANG .ASC) ANALYSIS

PITCH No.	fundamental frequ.(Hz)	pitch period(msec)	data pts
1	526.3	1.9	19.0
2	526.3	1.9	19.0
3	142.9	7.0	70.0
4	128.2	7.8	78.0
5	122.0	8.2	82.0
6	125.0	8.0	80.0
7	123.5	8.1	81.0
8	117.6	8.5	85.0
9	120.5	8.3	83.0
10	116.3	8.6	86.0
11	120.5	8.3	83.0
12	119.0	8.4	84.0
13	119.0	8.4	84.0
14	122.0	8.2	82.0

4. 합성단위 추출 및 스파이크성 잡음 제거

본 논문에서 음성합성 단위로 1 피치주기 구간의 음성파형을 추출하여 사용하였는데, 이는 한국어 무제한 규칙합성시 시간영역에서의 다른 합성단위 보다 규칙합성에 필요한 파라메타의 제어가 용이하기 때문이다. 그림 8(a)는 추출된 1 주기구간의 파형을 나타내며, 그림 6(c)는 추출된 1 주기의 파형을 연속적으로 나열시켜 합성시켰을 때의 스펙트럼도이다. 이 그림에서 스펙트럼 영역 전역에 일정한 간격으로 스파이크성과 같은 잡음이 오염된 현상이 발생되는데 이는 추출된 1 주기 파형의 시작점과 종료점에서

표 2-2. 음성합성용 단음절 데이타 정보 추출표
Table 2-2. Extracted information table for the synthesis of speech

SPEECH SIGNAL (LANG .ASC) ANALYSIS

No.	DATA POINT (PTS)	MAX (MV)	INTERUAL (PTS)	DATA POINT (PTS)	MIN (MV)	INTERUAL (PTS)	No. ZERO CROSSING	MAX RATE	MIN RATE
1	35	29	35	81	− 30	81	5	0.05	0.03
2	135	323	100	100	− 265	19	4	0.54	0.27
3	216	441	81	170	− 520	70	6	0.74	0.52
4	274	362	58	248	− 697	78	6	0.61	0.70
5	353	441	79	330	− 814	82	8	0.74	0.82
6	433	519	80	410	− 893	88	10	0.87	0.90
7	543	598	110	491	− 893	81	10	1.00	0.90
8	583	558	40	576	− 893	85	8	0.93	0.90
9	666	519	83	659	− 893	83	7	0.87	0.90
10	751	539	85	745	− 951	86	8	0.90	0.96
11	834	558	83	828	− 971	83	8	0.93	0.98
12	918	558	84	912	− 971	84	8	0.93	0.98
13	1001	558	83	996	− 971	84	8	0.93	0.98
14	1127	578	126	1078	− 991	82	10	0.97	1.00

의 파형 접합부에서의 불연속점에 기인된 것으로써 합성음의 음질을 저하시키는 주 요인의 하나이다. 스펙트럼 파형도가 잡음으로 오염되는 이유를 고찰하면, 추출된 1 주기 파형의 지속시간은 그림 8(a)에서와 같이 8.1msec 이며 이 때 8.1msec의 폭을 지닌 1 피치주기의 파형을 나열시키면 8.1msec 간격으로 파형 접합부에서 불연속 파형이 발생하는데 이를 주파수영역으로 변환시키면 1/8.1msec ≈ 123hz정도가 된다. 그림 6(a)에서와 같이 전체 주파수 폭은 5 khz 이므로 5000hz/123hz ≈ 41개 가량의 불연속점이 123 hz 간격으로 발생되며, 이를 표시한 그림이 6(c)와 6(d)이다. 추출시 이러한 불연속점이 최소가 되도록 추출하여야 한다. 본 논문에서는 표 2-2에서와 같은 파형 분석을 통하여 단음절내에서의 매 피치구간을

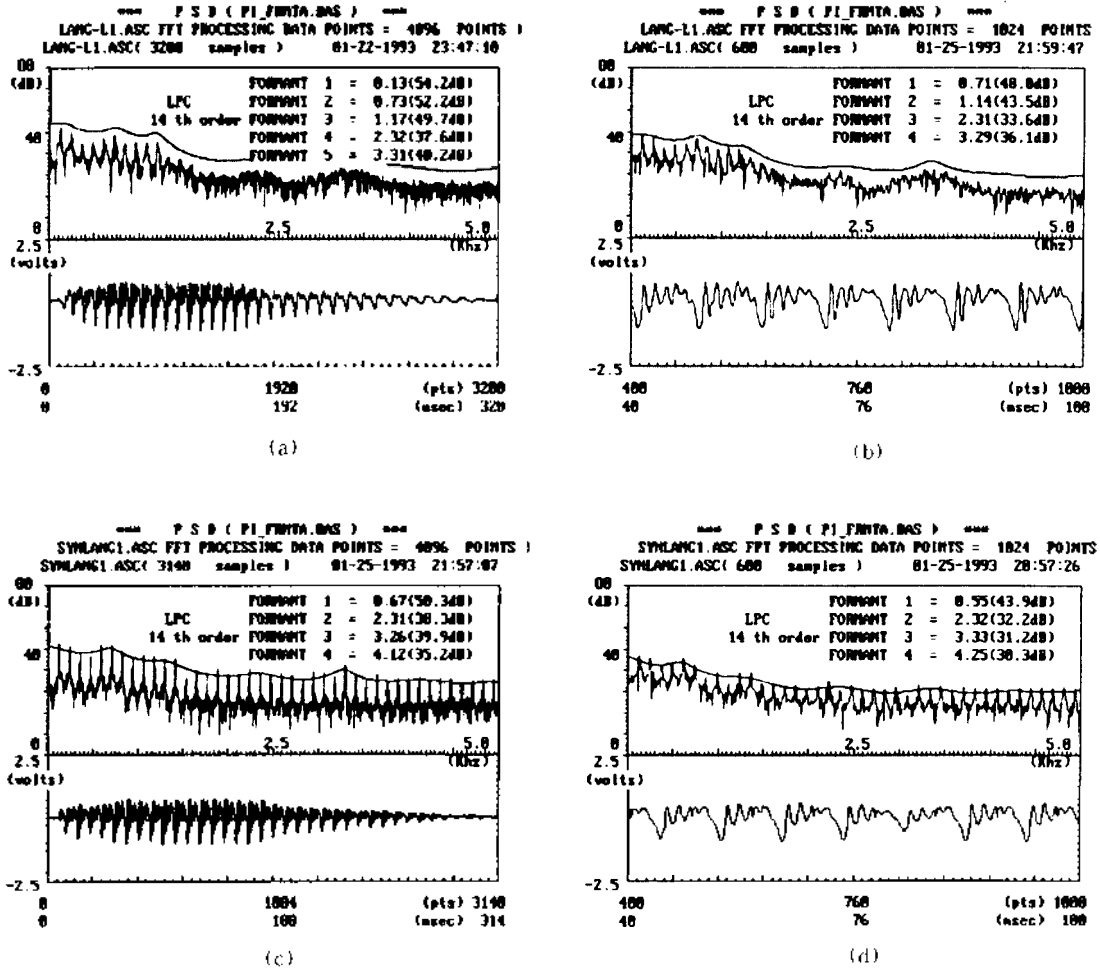


그림 6. (a)음성파형 "랑"의 스펙트럼
(b)그림 (a)의 분절파형 스펙트럼
(c)합성파형 "랑"의 스펙트럼
(d)그림 (c)의 분절파형 스펙트럼

Fig 6. (a)Spectrum of speech "lang"
(b)Spectrum of segmented speech in Fig 6(a)
(c)Spectrum of synthesized speech "lang"
(d)Spectrum of segmented speech in Fig 6(c)

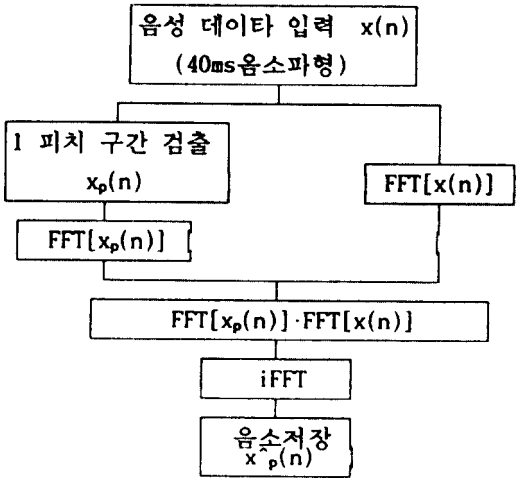


그림 7. 합성음 개선을 위한 디지탈 신호 처리
Fig 7. Digital signal processing for the improvement of synthesized speeches

검색하여 피치주기 간격에서의 파형의 정상부와 최저부를 구하여 이웃한 피치의 중심값을 취하여 그림 8(a)에서와 같은 임펄스 형태의 1 피치 구간을 추출하여 그림 5(b)에서 구한 기본 파형으로 피치를 동기시켜 파형을 합성하였다. 또한 그림 6(c)의 합성음에서 발생되는 문제점으로 1 피치주기 구간의 음성파형을 연속적으로 나열시키므로 합성음 청취시에는 조음변화 현상이 유발되지 않으므로 매 40msec 정도의 간격으로 피치를 추출하여 합성시킴으로써 음절내에서의 조음변화 현상을 유발시킬 수 있었다. 그러나 추출된 1 피치 프레임 단위의 음성 파형은 40msec 정도의 구간 내에서 추출된 파형이므로 이를 보상시키기 위한 처리과정이 그림 7의 디지탈신호처리 과정이다. 일정한 주기를 지닌 1피치 주기의 6내지 9ms 정도의 추출된 파형 $x_p(n)$에 40ms 정도의 음소파형 x(n)을 디지탈 필터링 처리(iFFT[FFT[$x_p(n)$]·FFT

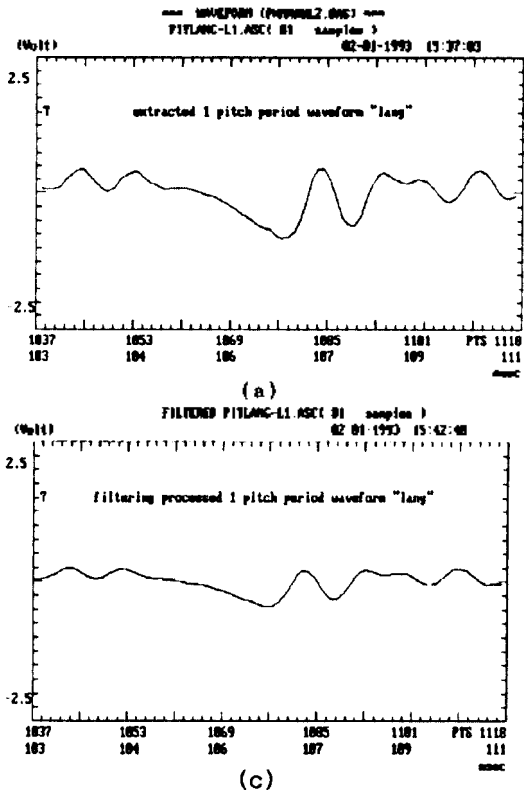
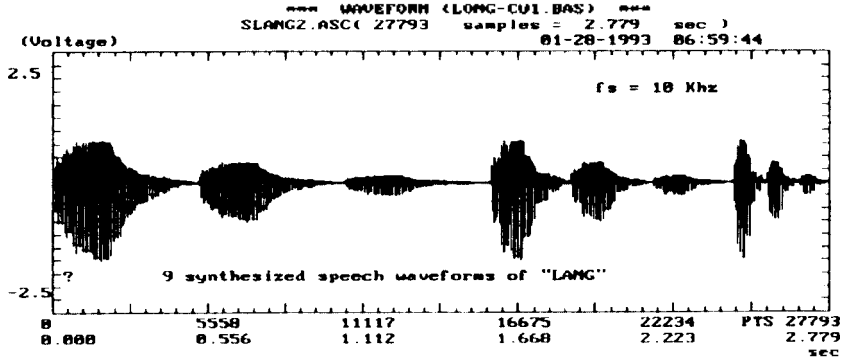


그림 8. (a)1 피치 구간의음성 파형 '랑'
(b)그림 (a)의 스펙트럼
Fig 8. (a) Waveform of 1 pitch period 'lang'
(b)Spectrum of Fig 8(a)
(c)디지탈 신호 처리된 1 피치 구간의 음성 파형
(d)그림 (c)의 스펙트럼
(c) Waveform filtered 1 pitch period 'lang'
(d)Spectrum of Fig 8(c)



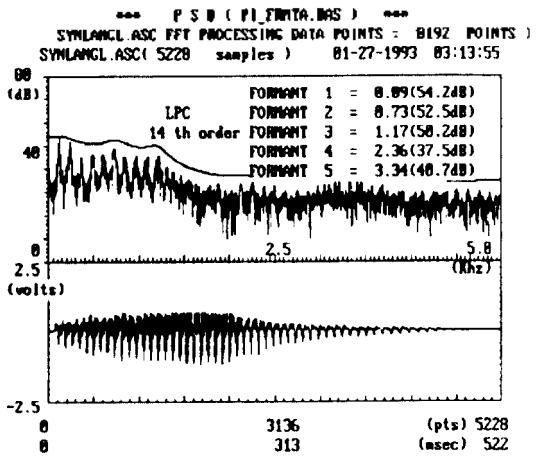
(a)

그림 9. (a)단음절 "랑"의 규칙합성 예
Fig 9. (a)Examples of synthesized by rule "lang"

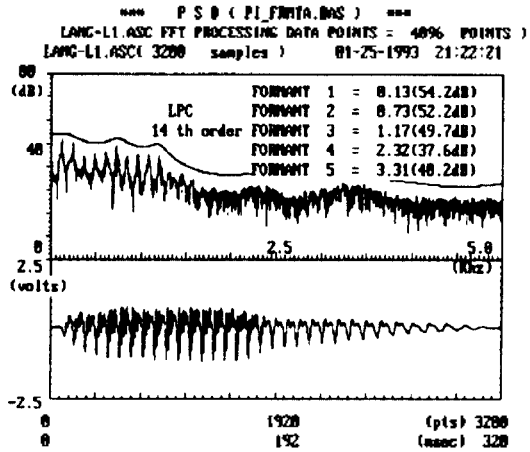
[x(n)]})시켜 구한 1 피치주기 구간의 파형 $\hat{x}_p(n)$이 그림 8(c) 이며, 이러한 처리과정을 거쳐 추출된 파형 $\hat{x}_p(n)$을 저장시켜 합성에 이용한 결과를 3장에 제시하였다.

Ⅲ. 합성 결과

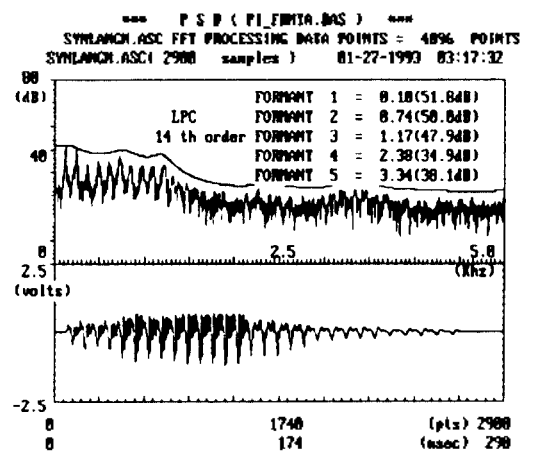
그림 7은 추출된 1 피치구간의 음성 파형을 이용하여 합성시켰을 때 발생되는 제반 문제점들을 제거시키기 위하여 사용된 디지탈 신호 처리 과정을 나타낸 것이다. 음성 신호는 시간에 따라 서서히 변하는 시변 함수(slowly time-variant system)이므로 단음절



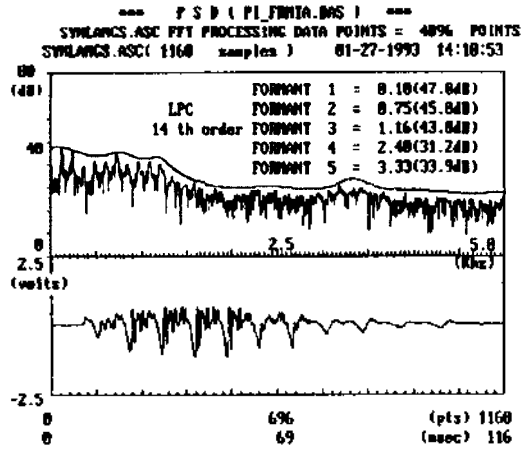
(b)



(a)



(c)

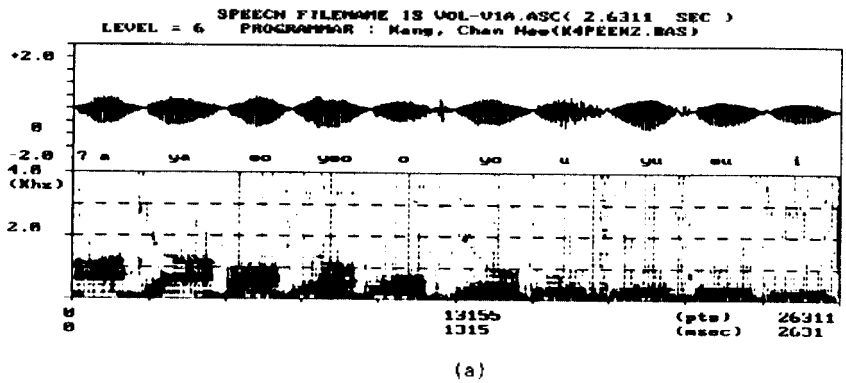


(d)

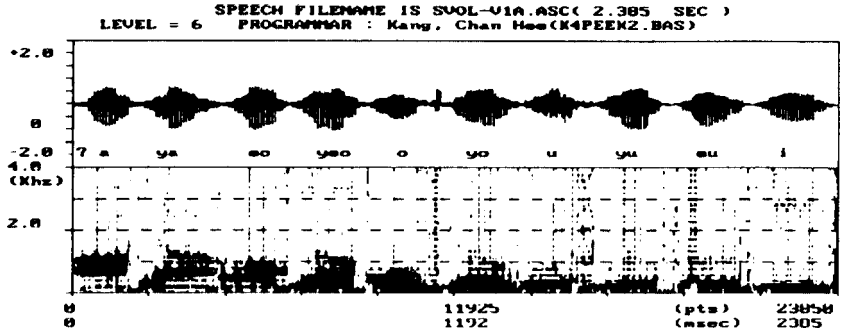
그림 10. (a)원음 "랑"의 스펙트럼
(b)규칙합성음(장음) "랑"의 스펙트럼
(c)규칙합성음(정상음) "랑"의 스펙트럼
(d)규칙합성음(단음) "랑"

Fig 10. (a)Spectrum of original speech "lang"
(b)Spectrum of synthesized long speech "lang"
(c)Spectrum of synthesized normal speech "lang"
(d)Spectrum of synthesized short speech "lang"

합성시 단시간 분석을 통한 합성 방법이 보편적으로 사용되고 있다. 따라서 본 논문에서도 1 음절어의 데이타를 획득한 후 40ms 정도의 음성 데이타로 부터 1 피치 구간의 음성 파형을 매 구간별로 추출하여 합성하였다. 이 때 발생되는 문제점은 추출된 1 피치구간의 음성파형에서와 같이 시작점과 종료점에서 다른 파형과 연결시켰을 때 불연속점이 발생하여 그림 6 (c)와 같은 형태의 합성음이 발생되어 음질이 매우 저하되는 현상이 초래되었다. 또한, 1 음절어 합성시 합성에 이용된 합성단위를 1 피치 구간의 음소파형을 1개 추출하여 연속적으로 연결시켜 합성시켰을 때는 1 음절내에서 조음기관이 서서히 변하는 조음변화 현상이 발생하지 않아서 합성음이 공명음화되는 현상이 발생하였다. 이러한 현상을 보정시키기 위하여 매 40 ms 구간내에서 1 피치 구간의 파형을 추출하여 합성시킴으로써 이러한 현상을 제거시킬 수 있었다. 그림 12는 이러한 합성방식에 의하여 규칙합성시킨 문장단위의 합성음에 대한 시간 파형 및 스펙트럼도를 나타낸 것이다. 그림 9는 저장된 음소화일과 합성용 매개변수를 사용하여 음성 "랑"을 여러가지 형태로 규칙합성시킨 예이며, 그림 10은 그림 9에서의 합성음과 원음을 스펙트럼 영역상에서 비교한 것이다. 그림 11은 한국어 단모음과 이중모음에 대한 합성 결과이다. 그림 9에서와 같이, 단음절 합성시 진폭과 지속시간의 변화가 가능하였으며, 음질에 대한 이해도와 명료도도 시간영역에서의 특징인 자연음에 가까운 결과를 얻었다. 그림 12와 13은 문장단위의 합성에 적용시키기 위한 가능성을 타진하기 위하여 극히 제한된 문장에 적용시켜 규칙합성시킨 결과를 제시한 것이다. 이 경우에는 단음절 합성에 대한 청취도 결과보다 훨씬 더 양호한 결과를 얻었다. 이는 문장 단위의 합성시에는 음절단위에 대한 합성음의 명료도 보다는 문장성분에 따른 진폭 패턴, 지속시간 패턴, 피치 패턴, 악센트 패턴과 쉼(pause) 패턴등의 파라메타가 문장 전체의 이해도, 명료도 및 자연성에 크게 영향을 미치게 되는 것을 확인하였다. 이 방식은 주파수 영역에서의 합성방식 보다 음질 및 자연성에 있어 우수하였다. 그 이유를 LP(선형예측 합성방식과 비교하여 설명하면 먼저, LP에 의한 방식에서는 단



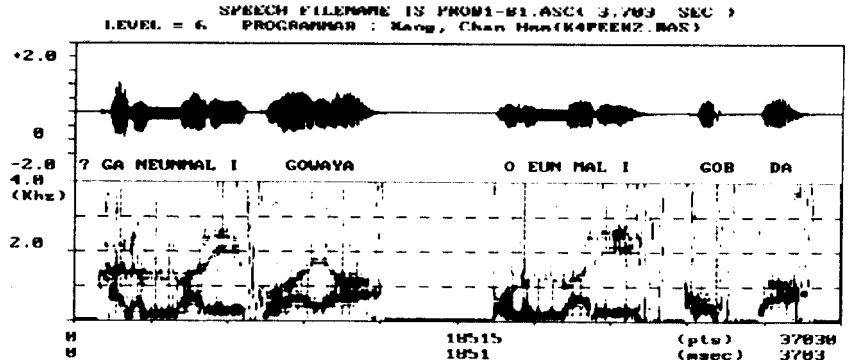
(a)



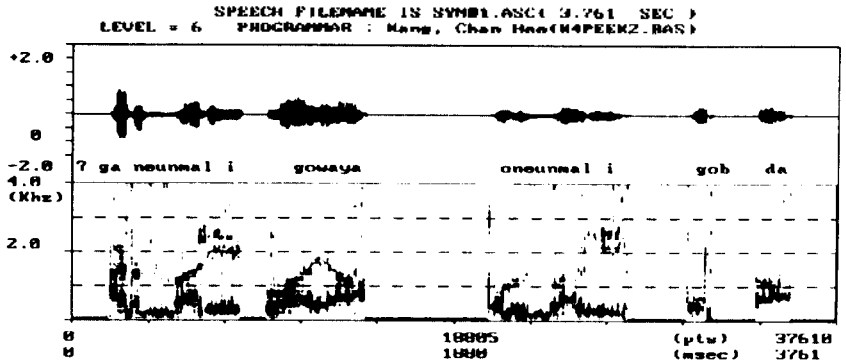
(b)

그림 11. (a)한국어 단모음과 이중모음의 원음 스펙토그램
("아, 야, 어, 여, 오, 요, 우, 유, 으, 이") (b)그림 (a)의 합성음 스펙토그램

Fig 11. (a)Spectrogram of original speech in Korean vowels(a, ya, eo, yeo, o, yo, u, yu, eu, i) (b)Spectrogram of Fig.11(a)



(a)



(b)

그림 12. (a)속담 "가는 말이 고와야 오는 말이 곱다."의 스펙토그램
(b)그림 (a)의 합성음에 대한 스펙토그램

Fig 12. (a)Spectrogram of Korean probe "ganeun mali gowaya oneun mali gob da."
(b)Spectrogram of synthesized sentences in Fig.12(a)

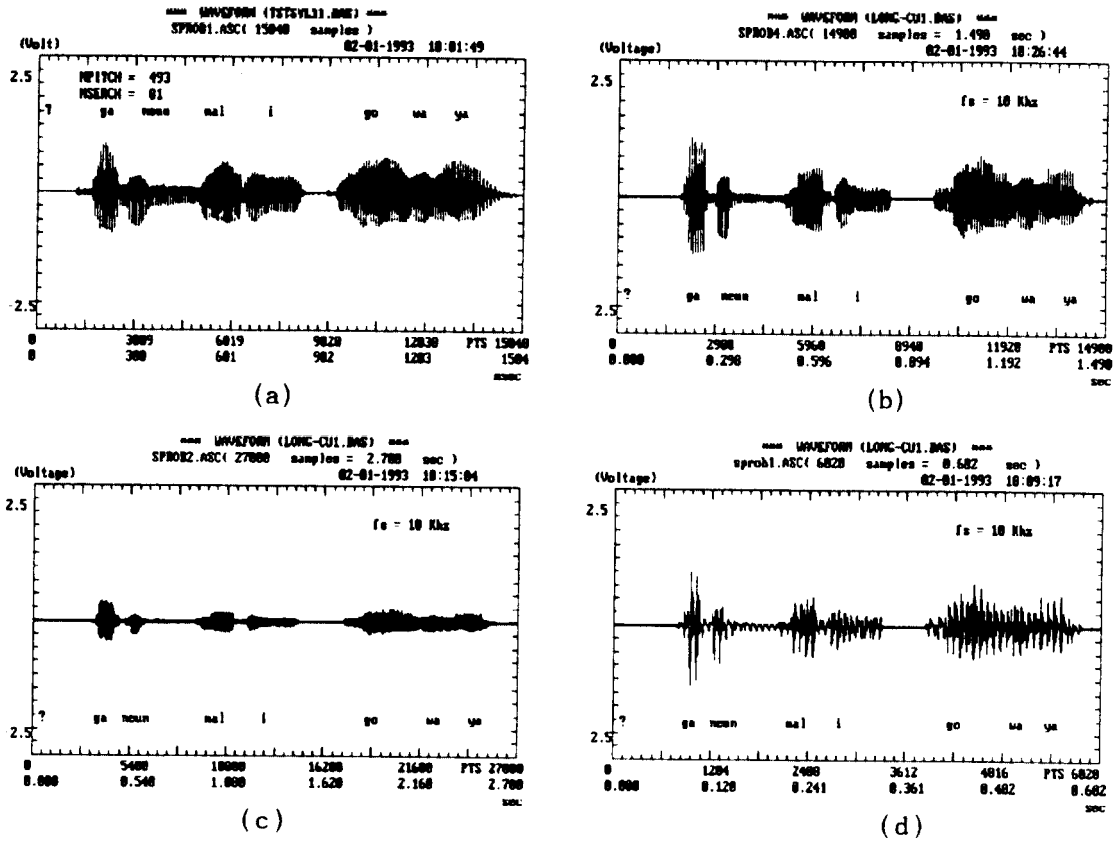


그림 13. (a)원음 파형 "가는 말이 고와야"
(b),(c),(d)규칙합성 예(장성음, 장·약음, 단음)

Fig 13. (a)Waveform of original speeches "ganeun mali gowaya"
(b)Examples of synthesized speeches(normal sound, long·weak sound and short sound)

시간(대략 25ms)간격으로 음성을 분석하여 음원을 생성시키기 위하여 무성음과 유성음을 판별하여 합성시키고자 하는 음성이 유성음일 경우에는 성대의 진동이 준주기성이므로 피치주기를 추정한 후 다시 음성의 에너지 정도를 판별하여 그에 비례하는 진폭을 지닌 준주기성 펄스를 발생시켜 음원을 구한다. 무성음일 경우에는 무질서한 잡음(random noise)원을 이용한 비주기성 잡음을 발생시켜 합성에 이용한다. 이와같이 LP방식에서는 유무성음 구간에서의 음원에 의하여 발생된 여기 신호가 성도에 인가되어 합성음을 발생시키는 방식이므로 음원이 합성음질에 중요한 역할을 하는 반면에 본 논문에서 제시한 방법에서는 음성파형을 직접 이용하는 방식이므로 음원과 성도 모델을 분리시킬 필요가 없으므로 음원에 의한 음질 저하 현상이 전혀 수반되지 않는다. 또한 LP방식에서의 성도모델은 LPC 계수를 추정하여 구한 것으로 라티스 필터등과 같은 구조로 필터를 구성시킨 후 앞서 구한 음원으로 필터링시켜 합성음을 발생시키는 데 반하여, 이 합성 방식은 시간영역에서의 합성방식이므로 그림 5(b)에서와 같이 정확하게 피치주기를 추출하여 피치 동기를 시킨 후 원 음성파형의 패턴에 매칭(PSPM : Pitch Synchronous Pattern Matching)시키는 합성방식이므로 피치제어가 매우 용이하며, LP에 의한 합성 방식에서는 성도 모델을 추정하여 단시간 구간내에서 연속적으로 합성음을 합성시키는데 반하여 이 방식에서는 정확한 성도모델을 추정한 것과 같이 실제의 음성파형을 1 피치 구간 추출하여 합성시키므로 주파수 영역에서의

합성방식과는 달리 수학적 모델링에 의한 오차가 발생하지 않아 합성음질이 우수하였으며 또한, 이 방식에서 사용한 바와같이 너무적은 양의 합성단위 사용으로 인하여 발생되는 음질 저하 요소를 제거시키기 위하여 디지탈 신호처리하여 구간별 음질 보정을 행하여 단구간내에서 연속적으로 합성시키므로 보다 더 음질이 우수한 합성음을 발생시킬 수 있었다.

Ⅳ. 결 론

본 논문에서는 음소단위에 의한 규칙합성시 수반되는 음질의 저하현상을 극복하기 위하여 사용된 처리기법을 컴퓨터 시뮬레이션으로 시행한 결과를 원음과 비교 제시하였다. 그 결과 음소단위에 의한 합성시 발생되는 파형의 접합부에서의 불일치로 인한 스펙트럼 왜곡현상을 제거시켜 음질의 향상을 이룰 수 있었으며, 음성합성시 피치를 정확하게 제어하여 발생시킬 수 있었다. 합성 결과 미흡한 점으로는 음성합성에 사용되는 한국어의 규칙을 여러가지로 더 세분화시켜 합성시키면 보다 더 양호한 음질을 보완시킬 수 있을 것이다. 아직도 개선의 여지가 많이 남아 있기는 하지만 한국어 문어전환 시스템내에서의 음성합성기로 사용되기 위하여는 자연언어처리부의 개발이 시급히 요구된다. 지금까지의 연구 결과로는 운율요소중 강음과 단음의 발생은 이 방식이 저장된 1 피치주기 단위의 파형을 직접 처상하여 연속적으로 연결시켜 합성시키는 방식이므로 음절단위의 지속시간을 조정함으로써 장단음의 제어가 가능하였으며, 강약과 같은 요소는 저장된 음소파형의 진폭을 악센트 성분에 따라 적당 비율로 조절함으로써 강약조절을 이룰 수 있었다. 억양은 피치주기에 의하여 이루어진다는 사실은 이미 널리 알려져 있으며 문어전환 시스템내에서의 음성합성기 설계에 있어서 매우 중요한 연구과제이다. 특히 문장단위에 의한 합성시 피치 패턴에 관한 연구가 이루어져야 할 것이다. 이는 자연성에 가장 큰 영향을 미치는 요소로써 한국어 문장에 대한 피치 패턴의 데이타 베이스화 작업이 이루어져야 할 것이다. 이러한 결과를 이용하여 문장단위를 합성할 경우 여기서 사용된 방식에서는 원도윙 처리[10]를 이용하여 피치패턴의 제어가 가능할 것으로 추정되어지나 이에 대한 연구는 앞으로 계속 이루어져야 할 것이다. 또한, 음질의 향상을 기하기 위하여는 2 음절이 이상의 단어, 구문, 문장에 대한 한국어가 지니고 있는 음운학적 특징 추출 및 정확한 파라메타의 분석에 대한 단계적이고도 체계적인 연구가 이루어져야 한다. 마지막으로 한국어의 음소단위에 의한 규칙합성시 음질단위의 고립어 음질은 개선되었으나, 문장단위의 부자연한 합성을 실현하기 위하여는 문장을 구문별로 분석, 처상, 분석하여 억양, 강세, 장단 등과 같은 운율정보를 부여하기 위한 단어 사전을 데이타 베이스화 하여야 하며, 문장 전체적인 흐름을 자연스럽게 처리할 수 있도록 한국어 문장별 피치 패턴 발생을 위한 알고리즘 개발이 이루어져야 할 것이다.

참 고 문 헌

1. 은종관, 진효섭, 김형용, "무제한 한국어 합성에 관하여," 과학기술처 '87 특정연구진과발표회 논문집 88. 7
2. 강화대 멀티 정보 시스템 공학 연구소, "CYBEX 시스템－메세지 페이저를 중심으로 ," 1992
3. Shuzo Saito, Speech Science and Technology, IOS press, 1992
4. Jhonathan Allen, M. Sharon Hunnicutt, Dennis Klatt, From text to speech : The MITalk System, Cambridge University Press, 1987
5. Shuzo Saito, Fundamentals of Speech Signal Processing, Academic Press, 1981
6. Geoff Bristow, Electronic Speech Synthesis : Techniques, Technology and Applications, McGraw-Hill Book Company, 1984
7. 윤선규, 강찬희, 진용옥, "한국어 자모음간의 천이영역 구분을 위한 Ambiguity 패턴에 관한 연구," 한국음향학회 학술대회 논문집, 1988. 10
8. 강찬희, 김병용, 진용옥, "단위피치 프레임 정보를 이용한 한국어 음질 합성," 한국통신학회 학술대회 논문집, 1989. 8
9. 강찬희, 진용옥, "동계적 신호처리에 의한 자모형 한국어의 합성," 1990. 5 한국통신학회 춘계학술 발표회 논문집, 1990.5
10. Hamon, C., Moulines, E. & Charpentier, F., "A diphone synthesis system based on time-domain modification of speech," Procedings of ICASSP 89, 238-241

▲姜 贊 熙(Chan Hee Kang) 1958년 3월 6일생

1980년 2월 : 경희대학교 전자공학과 졸업
1982년 2월 : 경희대학교 공학석사
1982년 7월 : 해군장교 임관
1983년 7월 ~ 1985년 7월 : 해군사관학교 교수부 전자과
1985년 8월 ~ 1986년 8월 : 삼성반도체 통신연구소
1986년 9월 ~ 1989년 8월 : 경희대학교 박사과정 수료
1989년 3월 ~ 현재 : 상지대학교 병설 전문대학 전자과

▲陳 庸 玉(Yong Ohk Chin) 1943년 3월 21일생
1968년 2월 : 연세대학교 공과대학 전기공학과 졸업
1975년 2월 : 연세대학교 대학원 전자공학과(공학석사)
1981년 8월 : 연세대학교 대학원 전자공학과(공학박사)
1980년 : 통신기술사
1976년 ~ 현재 : 경희대학교 공과대학 전자공학과 교수