論文98-35C-7-7

시간 영역에서의 무제한 고립어 합성을 위한 운율 요소 제어용 알고리즘 개발

(Development of An Algorithm for the Control of Prosodic Factors to Synthesize Unlimited Isolated Words in the Time Domain)

姜 贊 熙 *

(Chan Hee Kang)

요 약

본 논문은 한국어 무제한 규칙 합성을 위한 합성 알고리즘 개발에 관한 연구이다. 시간 영역상에서 고립어 단위를 기본 합성음으로 사용하여 운율요소를 제어한 결과를 제시한다. 본 논문에서 제안한 합성 방식은 pitch-synchronous하고 parametric한 시간영역 상에서의 합성방식으로서, 여기서는 주로 피치 주기 및 에너지 윤곽선 그리고 지속시간 등과 같은 운율요소의 제어와 합성음에 대한 음질평가 결과를 제시한다. 합성시, 3가지 운율요소의 연속적인 통합제어에 의한 연속어 합성이 가능하여져 자연성이 향상되었다. 실험 결과, 파형 접합 면에서의 피치 불연속성과 에너지 제로점 등이 개선되었다(그림 6). 특히, 매우 다양한 지속시간과 음조 패턴을 지닌 합성음을 거의 무제한 적으로 합성 가능하였다(그림 5, 그림 6). 파형 접합 점에서의 불일치로 인한 위상왜곡에 의한 음질 저하 현상도 개선되어 잡음감 및 명료도도 양호하였다(표1, 그림 7).

Abstract

This paper is to develop an algorithm for the unlimited Korean speech synthesis. We present the results controlled of prosodic factors with isolated words as a synthesis basis unit in the time domain. With a new pitch-synchronous and parametric speech synthesis method in the time domain here we mainly present the results of controlled prosody factors such as pitch periods, energy envelops and durations and the evaluation of synthetic speech qualities. In the case of synthesis, it is possible to synthesize connected words by controlling of a continuous unified prosody that makes to improve the naturalities. In the results of experiment, it also has been to be improved uncontinuities of pitch and zeroing of energy in the juncton parts of speech waveforms(fig.6). Specially it has been to be possible to synthesize speeches with unlimitted durations and tones(fig.5, fig.6). So on it makes the noisiness and the clearness better by improving the degradation effects from the phase distortion due to the discontinuities in the waveform connection parts(table 1, fig.7).

* 終身會員, 尙志大學校 併設 專門大學 電子科

(Dept. of Electronic Eng., Sangji Jr. Coll.)

※ 이 논문은 1997년도 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음.

接受日字:1998年1月22日, 수정완료일:1998年6月16日

Ⅰ. 서론

본 논문은 한국어 TTS 시스템내 새로운 음성 합성 알고리즘 개발을 위한 1차 연구 결과이다. 규칙합성음의 음질저하를 극복하고 운율요소의 제어가 용이하도록 추출된 파라미터를 합성에 이용함으로써 자연성을

개선시키기 위한 시간 영역에서 무제한 한국어 음성합성방식의 알고리즘을 제안한다. 이에 대한 타당성 검토로써 음질 평가를 수행한 결과를 제시한다. 새로운 음성 합성 방식의 개발에 있어서 무엇 보다 중요한 점은 운율요소의 제어이다. 이 때 합성음의 음질 또한 자연스럽고 명료하여야 한다. 운율요소의 제어에는 피치 주기, 에너지 윤곽선 및 지속시간의 제어 등을 들 수 있다. 첫째로, 피치 주기의 제어란 음성 합성 시 억양의 제어를 의미하며, 합성음의 리듬감을 나타내는 중요한 요소가 된다. 따라서 구문분석부에서 입력된 문장을 분석하여 수많은 형태의 억양을 그 때의 문맥 상황에 따라서 생성시킬 수 있어야 한다. 둘째로 에너지 윤곽선의 제어는 합성음의 강약 성분을 제어하는 것을 의미하며, 다양한 형태의 에너지 윤곽선을 연속적으로 생성할 수 있도록 제어하여야 한다. 마지막으로 음성의 지속시간 제어란 음의 장단을 제어하는 것을 의미한다[1),-3)]. 명료하고 자연스러운 합성음을 생성하려면 이들 세 가지 형태의 운율요소들을 하나로 통합시켜 규칙 합성음을 생성하여야 한다.

일반적으로 주파수영역에서의 합성방식은 음원 부와 성도 부를 추정하여 규칙 합성한다. 이 때 규칙합성용 매개변수에 의하여 운율요소를 제어하여 규칙 합성음을 생성시킨다. 따라서 운율요소의 제어가 용이하여 자연성에 있어서는 시간 영역에서의 합성 방식 보다 우수하다. 그러나 성도 부와 음원 부에서 발생된 추정 오차로 인하여 음질에 있어서는 명료성이 떨어진다[4)-6)]. 이러한 단점을 보완한 것이 TD-PSOLA방식이다[7)-8)]. 이 방식은 non-parametric한 방식으로서 CDU(context dependent text) 단위로 분할하여 피치정보와 음가정보와 에너지 윤곽선 정보를 저장시킨 후, 규칙에 따라 합성하는 방식이다. 피치주기의 변경은 pitch-synchronous하게 윈도우 함수를 가하여 피치 주기를 짧게 하든가 길게 변경시킴으로서 억양 성분을 제어한다. 지속시간의 변경은 ST(short-time)를 적당히 반복하거나 생략시킴으로서 지속시간을 제어한다. 또한 에너지 윤곽선의 제어는 저장된 CDU의 마지막 에너지와 연결할 CDU의 첫 부분의 에너지 차를 선형적으로 연결시켜 제어하는 시간영역에서의 합성 방식이다.

본 논문에서 제안한 방식은 TD-PSOLA 방식과는 달리, 기본 합성 단위음으로 사용한 고립어로부터 추출된 운율제어 정보들을 매개변수화하여 시간영역 상에서 규칙 합성시키는 방식이다. 이는 매개변수의 추출에 의한 parametric한 방식의 일종이므로 주파수영역에서의 운율요소 제어방식과 개념적으로는 유사하다. 그러나 저장된 음성 데이터를 사용하여 시간영역에서 합성하는 방식이므로 TD-PSOLA 방식이 지니고 있는 non-parametric한 방식의 일종이다. 따라서 본 논문에서는 주파수 영역에서의 합성방식이 지니고 있는 운율요소 제어의 용이성으로 인한 우수한 자연성과 시간영역 상에서 합성방식이 지니고 있는 음질의 양호성을 모두 지닌 합성방식을 개발하고자 하였다. 또한, 합성단위로 다이폰을 사용할 경우에는 다이폰 단위의 접속 시, 에너지 차 등으로 인한 에코 현상이 수반되므로 본 논문에서는 이러한 현상을 최소화하기 위하여 고립어를 기본 합성 단위음으로 사용하였다.

일반적으로 파형 연결에 의한 합성 방법은 위상왜곡으로 인한 잡음이라든가, 접속 점에서의 불일치로 인한 스파이크성 잡음이 발생되어 음질이 저하되고 운율요소의 제어가 용이하지 못하다. 따라서 제안한 방식에서는 이러한 점들을 극복하기 위하여 운율요소를 시간영역 상에서 효율적으로 제어하기 위한 몇 가지 매개변수에 의한 합성 방법 및 그 결과를 제시한다. 이를 위하여 본 논문에서는 고립어 단위를 사용하였을 때 발생되는 문제점을 극복하기 위하여 저장된 합성단위음으로부터 강약, 장단 및 고저성분을 자유로이 합성시킬 수 있는 시간영역에서의 운율요소의 제어가 가능한 알고리즘 개발에 대하여 주로 논한다. 즉, 1)지속시간(장단요소) 제어 알고리즘 개발을 위하여 200내지 300ms 정도로 구축된 단음절어 단위의 합성단위의 합성음을 사용하여 한국어 4가지 유형의 모든 고립어를 80msec 정도 부터 1sec 사이의 지속시간을 지닌 합성음을 자유자재로 제어 가능한 시간영역에서의 합성 알고리즘을 개발하고 그 결과를 제시한다(표 1, 그림 7(b)). 2)진폭(강약요소) 제어 알고리즘 개발을 위하여는 저장된 고립어 합성단위음의 특성상 시작점과 끝점의 에너지 패턴이 영이 되어 연속음 합성시킬 경우 음절 접속 구간에서 끊어지는 듯한 외마디음의 합성음(그림 3)과 음절간 불협화음으로 인한 에코현상이 발생되는 것을 극복하기 위하여 한국어 4가지 유형에 대한 모든 고립어의 진폭 포락선 형태에 대한 유무성음 구간 및 피치구간별 진폭 포락선 형태를 일정 비율로 추출한 진폭 제어용 파라메타를 이용하여 합성시킨 결과를 제시한다(그림 6). 마지막으로 피치(억양요소)

제어 알고리즘 개발을 위하여는 고정된 단위합성음의 억양으로부터 규칙합성 과정 중에 합성하고자 하는 임의의 억양 패턴을 자유자재로 합성 가능한 피치 제어 알고리즘을 제안하고 그 결과를 제시한다(그림 5, 그림 7(c)). 본 논문은 한국어 TTS 시스템내 음성합성부를 개발하기 위한 1차 연구과제로써 무제한 고립어 합성을 위한 운율요소 제어 알고리즘 개발로 국한하였으며, 다음절어 및 문장단위의 무제한 규칙 합성시 발생되는 문제점 즉 음절간 결합시 조음변화에 따른 파형 포락선 형태 제어, 다음절어 규칙합성을 위한 음운기호 설계, 다음절어 및 문장단위 연속음 규칙합성을 위한 강약규칙, 지속시간 패턴 및 억양패턴의 DB구축 등 수 많은 과제는 향후로 미룬다.

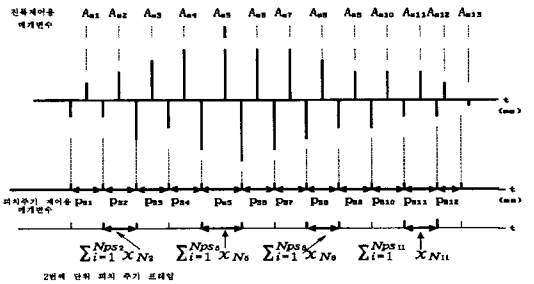


그림 1. 음성 파형 분석에 의한 운율 제어용 매개변수 추출

Fig. 1. The extraction of prosody control parameters by a speech waveform analysis.

Ⅱ. 규칙 합성용 매개변수 및 운율요소 제어

본 논문에서는 고립어 단위의 음성 파형을 입력시켜 파형을 분석한 후, 규칙 합성용 매개변수를 추출하여 지속시간, 강약 및 억양(피치 주기)등과 같은 운율요소를 시간 영역 상에서 제어하여 합성하였다. 주파수영역에서의 합성 시, 억양성분(피치 주기)은 임펄스 열 간격, 강약성분은 임펄스 세기, 장단성분은 프레임 개수 등을 조절하여 운율요소를 제어한다. 본 논문에서도 파형을 분석하여 이에 대응하는 운율 제어용 매개변수를 시간영역 상에서 추출하여 운율요소를 제어하는 방법을 제안하였으며, 사용된 주요 5 가지 매개변수는 다음과 같다.

1. 규칙합성용 매개변수

① 피치주기 제어용 매개변수 : 1PERIOD(고립어내 1 피치 주기 프레임 당 음성 데이터)

고립어 내 유성음 구간에서의 단위 피치 주기 프레임 총 갯수(변수:Np)와 단위 피치 주기 프레임 내 음성 데이터(변수:1PERIOD)에 3차 Lagrange 보간법을 적용시켜 변경시키고자 하는 새로운 피치 주기의 음성 데이터 열로 바꾸어 피치 주기를 제어함.

② 진폭 포락선 제어용 매개변수 : RATIO

식 2)에서와 같이 고립어 내 단위 피치 주기 프레임 간격 당 최대•최소 진폭 열을 추출하여 포락선형태를 제어함.

③ 지속시간 제어용 매개변수 : Np

단위 피치 주기 프레임의 총 갯수(변수:Np)를 조절하여 지속시간을 제어함.

④ 강약 제어용 매개변수 : MAX, MIN

식 2)에서와 같이 단위 피치 주기 프레임 간격 내에서의 최대•최소 진폭 비율을 조절하여 제어함.

⑤ 쉼 제어용 매개변수 : ZERO

구문 성격에 따라 0.2초, 0.3 초, 0.5초, 1초 등 적당한 쉼 구간 데이터를 부여하여 제어함.

먼저, 음성 파형으로부터 규칙 합성용 매개변수를 추출하기 위하여, 임의의 음성 데이터를 $\chi(n)$, 고립어의 음성 데이터 총 갯수를 N, 무성음부의 음성 데이터 갯수를 N_c, 유성음부의 음성 데이터 갯수를 N_v, 고립어 유성음부 내에서의 단위 피치 주기 프레임 총 갯수를 N_p로 각각 정의한다. 그러면, 고립어 내의 음성 데이터 열은 $\sum_{n=1}^{N} \chi(n)$ 으로 표기된다. 이 때, 각각의 단위 피치주기 프레임 당 구간의 경계를 P_{s1}, P_{s2}, P_{s3}, ···· 등으로 나타내고, 각 단위 피치주기 프레임 구간에서의 음성 데이터 갯수를 N_{ps1}, N_{ps2}, N_{ps3}, ···· 등을 배열 $N_{ps}(\ )$로 표기하면, N 개의 음성 데이터 열 $\sum_{n=1}^{N} \chi(n)$을 1 차원 배열인 단위 피치 주기 프레임의 N_p개 소 블록의 합으로 표기 가능하다. 따라서 이를 2차원 배열로 표시하면,

$$\sum_{n=1}^{N} \chi(n) = \sum_{i=1}^{Nc} x(i) + \sum_{j=1}^{Nv} x(j) \quad (1)$$

(단, $N = N_c + N_v$임.)

$$\sum_{j=1}^{Nv} x(j) = \sum_{n_1=1}^{Np} \sum_{n_2=1}^{Nps(i)} \chi(n_1, n_2) \quad (2)$$

이 된다. 여기서, $\chi(5,10)$는 5번째 단위 피치주기 프레임 구간의 10번째 데이터를 의미한다. 즉, $\chi(n_1,n_2)$는 χ(고립어 내 단위 피치주기 프레임 번호, 단위 피치 주기 프레임 구간 내 음성 데이터 열 번호)를 의미

한다. 또한, 단위 피치 주기 프레임 구간 내에서의 음성 데이터 열의 최대 진폭의 절대치를 각각 A_{m1}, A_{m2}, A_{m3}, ···· 등으로 정의하고, 각각의 피치 주기 프레임 구간내의 데이터 열을 일정한 크기로 정규 화 시킨 임의의 음성 데이터를 $\chi_N(n)$으로 정의하면, 2 차원 블록화 배열로 표시된 음성 데이터 $\sum_{n_1=1}^{Np}\sum_{n_2=1}^{Npa(i)} \chi(n_1, n_2)$은

$$\sum_{n_1=1}^{Np}\sum_{n_2=1}^{Npa(i)} \chi(n_1, n_2) \approx \sum_{n_1=1}^{Np}\sum_{n_2=1}^{Npa(n_1)} A_m(n_1)\cdot \chi_M(n_1, n_2) \quad (3)$$

이 된다. 위 식들로부터 추출·저장하여 고립어 단위 음성 DB에 작성된 추출된 주요 매개변수는

1) 고립어 내 전체 데이터 갯수 정보 : N(2 바이트)

2) 고립어 내 단위 피치 주기 프레임 갯수 정보 : N_p(1 바이트)

3) 각 피치주기 프레임 구간 내 음성 데이터 열 갯수 정보 : $\sum_{i=1}^{Np} N_{psi}$ (N_p 바이트)

4) 고립어 내 단위 피치 주기 프레임 당 최대진폭 정보 : $\sum_{i=1}^{Np} Am(i)$ (N_p바이트)

5) 단위 피치 주기별로 정규화된 음성 데이터 정보 : $\sum_{n_1=1}^{Np}\sum_{n_2=1}^{Npa(n1)} \chi_M(n_1, n_2)$ (N 바이트)

등이다.

2. 선형 보간법에 의한 피치 주기 변경

1) 라그랑쥐 보간법

보간법이란, $\chi_0, \chi_1, \ldots\ldots, \chi_n$ 들을 임의의 분점이라 하고, 이들 분점들이 이루는 최소구간을 $[\chi_0, \chi_n]$ 이라 했을 때 ,이 구간 내에 속하는 $\chi \in [\chi_0, \chi_n]$ 에 대한 $f(\chi)$를 추정하는 방법이다. 고차 Lagrange 보간법은 선형보간법의 일종으로서, n + 1 개의 상이한 분점 $\chi_0, \chi_1, \ldots\ldots, \chi_n$에서 실 변수함수 f (χ)의 함수치 $f(\chi_i)$ $(0 \le i \le n)$ 가 알려져 있을 때

$$f(\chi_i) = p(\chi_i) \; ; \; I = 0, 1, 2, \ldots\ldots, n \quad (4)$$

을 만족하는 n차 선형 보간 다항식 $p(\chi)$는 다음 성질을 갖는 n차 라그랑쥐 다항식, $L_{n,k}(\chi)$를 응용하여 $p(\chi)$를 보다 더 간단히 구할 수 있다.

$$L_{n,k}(\chi_i) = \delta_{ki} = \begin{cases} i = k \text{ 이면 } 1 \\ i \neq k \text{ 이면 } 0 \end{cases} \quad (5)$$

먼저, 구하고자 하는 $p(\chi)$를

$$\begin{aligned} f(\chi) \simeq p(\chi) &= L_{n,0}(\chi)\cdot f(\chi_0) + L_{n,1}(\chi) \\ &\quad \cdot f(\chi_1) + \ldots + L_{n,n}(\chi)\cdot f(\chi_n) \\ &= \sum_{k=0} L_{n,k}(\chi)\cdot f(\chi_k) \end{aligned} \quad (6)$$

와 같이 n차 라그랑쥐 다항식의 일차 결합으로 된다. $L_{n,k}(\chi)$이 n차 다항식이기 때문에 $p(\chi)$는 n차 다항식이다. 그리고 $L_{n,k}(\chi)$의 성질 때문에

$$\begin{aligned} p(\chi_i) &= \sum_{k=0} L_{n,k}(\chi_i)\cdot f(\chi_k) \\ &= \sum_{k=0} \delta_{ki}\cdot f(\chi_k) = f(\chi_i) \end{aligned} \quad (7)$$

이 되고, 이에 따라 $p(\chi)$가 식(4)의 조건을 만족한다. 따라서 식(6)의 $p(\chi)$는 n차 보간다항식이 된다.

식(5)에서 $L_{n,k}(\chi_i) = 0 (k \neq i)$이기 위해서는 n차 다항식 $L_{n,k}(\chi)$는 $\chi - \chi_i (i \neq k)$의 인수를 가지고 있어야 한다. 그러므로 $L_{n,k}(\chi)$는

$$\begin{aligned} L_{n,k}(\chi) &= C\cdot(\chi-\chi_0)\cdot(\chi-\chi_1)\cdot\ldots \\ &\quad \cdot(\chi-\chi_{k-1})\cdot\ldots\cdot(\chi-\chi_n) \\ &= C\cdot \prod_{(i=0, i\neq k)}^{n}(\chi-\chi_i) \end{aligned} \quad (8)$$

와 같이 $(\chi - \chi_i)(i \neq k)$의 인수의 곱으로 나타낼 수 있다. 단, 여기서 C는 상수인데 이는 $L_{n,k}(\chi_k) = 1$ 을 이용하면

$$1 = C\cdot \prod_{(i=0, i\neq k)}^{n}(\chi_k-\chi_i) \quad (9)$$

$$C = 1 / \prod_{(i=0, i\neq k)}^{n}(\chi_k-\chi_i)$$

됨이 알 수 있다. 따라서 $L_{n,k}(\chi)$는

$$\begin{aligned} L_{n,k}(\chi) &= \prod_{(i=0, i\neq k)}^{n}(\chi-\chi_i) \Big/ \prod_{(i=0, i\neq k)}^{n}(\chi_k-\chi_i) \\ &= \prod_{(i=0, i\neq k)}^{n}(\chi-\chi_i)/(\chi_k-\chi_i) \end{aligned} \quad (10)$$

으로 주어지며, 식(6)과 결합하면 n 차 라그랑쥐 보간 다항식 $p(\chi)$를 구할 수 있다.

2) 피치 주기 변경

언어에는 자음과 모음 같은 말소리 이외에도 이들에 얹혀서 나는 요소들이 있다. 이를 운율적 요소라 한다. 운율적 요소에는 소리의 길이를 나타내는 장단요소와 소리의 높이를 나타내는 고저요소(피치), 소리의 세기를 나타내는 강약요소등 3가지가 있다. 본 논문에서

피치주기의 변경이라 함은 말소리의 높낮이를 달리하여 합성하기 위하여 저장된 소리의 높낮이를 조절하여 새로운 소리의 높낮이로 변경시키는 것을 의미한다. 일반적으로 소리의 높낮이는 성대의 진동수에 따라 달라지는데, 진동수가 많을수록 소리는 높고 반대로 진동수가 낮을수록 소리는 낮다. 성대 진동에 따라 준주기성(pseudo-periodic)의 피치주기를 지닌 음성 파형이 생성되며, 성대가 1회 울릴 때 마다 1개의 준주기성 피치구간이 형성된다. 시간영역 상에서 피치 주기 변경에 있어서 문제점은, 피치 주기의 경계구간을 설정하여 단위 피치주기의 프레임을 결정짓는 것과 저장된 원음의 파형 형태를 그대로 유지한 채, 설정된 단위 피치 주기를 새로운 단위 피치 주기로 변경하는 문제다. 첫 번째 문제의 중요성은 서론에서 언급한 바와 같이 부정확한 피치 주기 열의 추정은 위상 왜곡과 파형 접합 면에서의 잡음을 초래하여 합성음의 음질을 저하시키는 주 요인이 되며, 두 번째 문제에 있어서는 기본 주파수가 120 Hz 인 피치 주기 열($83=1/f_0 * f_s = 1/120 * 10000$ if f_s = 10 kHz)을 기본 주파수가 80Hz인 피치 주기 열(125 개의 음성 데이터)로 원음이 지닌 파형의 형태를 유지한 채 변경시켜야 하는 점이다. TD-PSOLA 방식에서는 피치 주기를 변경시키기 위하여 pitch- synchronous하게 창 함수를 가하여 변경시키나, 창 함수의 영향으로 인하여 합성 시 다소 음질을 저하시키는 요인이 된다.

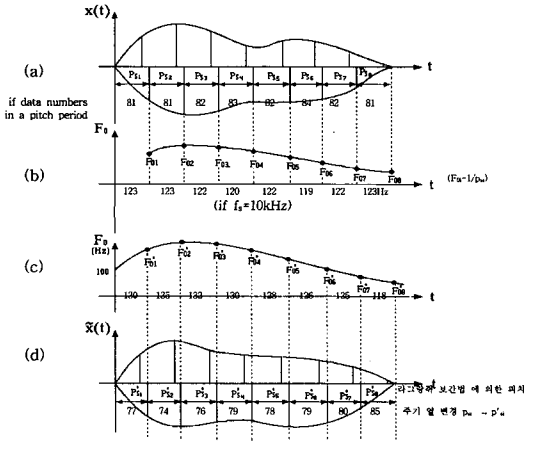


그림 2. 선형보간법을 이용한 피치 주기 변경
Fig. 2. The modification of pitch periods using a linear interpolation method.

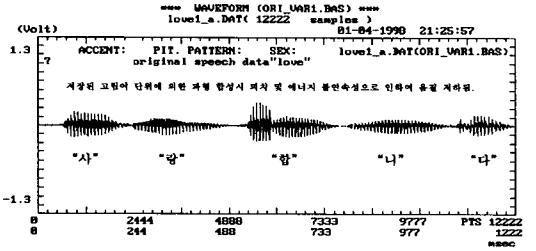
따라서 본 논문에서는 피치주기 변경 시, 원 음성의 음질에 변화를 초래하지 않고 합성할 수 있도록 선형 보간법을 사용하였다.

그 원리를 개략적으로 설명하면, 먼저 Ⅱ장 1절의 식(2)에서와 같이, 유성음 구간의 음성 데이터를 N_p개의 단위 피치주기 프레임 열로 분할저장시킨다. 그림 2(b)는 저장된 음성의 피치를 그림으로 나타낸 것이다. 식(2)에서 N_{ps}는 단위 피치 주기 프레임의 음성데이터 갯수이며, 그림 2(b)에서와 같이 기본 주파수의 산출은 $1 / N_{ps} * $ 주파수 샘플링 비(f_s)가 된다(1 피치 프레임 데이터 갯수가 80개 이면 기본주파수는 $1 / 80 * 10000\text{Hz} = 125\text{Hz}$로 됨). 그림 2(c)는 변경시키고자 하는 임의의 피치주기 패턴이며, 합성시키고자 하는 합성음의 지속시간에 따라 단위 피치주기 프레임 갯수를 결정한다. 그 다음에 각각의 단위 피치 주기 프레임에 라그랑쥐 보간법을 사용하여 변경시키고자 하는 피치주기(그림 2(c))를 지닌 새로운 음성 데이터 열(그림 2(d))로 변환시킴으로써 피치주기 변경 과정은 종료된다. 예를 들면, 그림 2(a)에서와 같이 저장된 음성파형의 파형분석과정에서 분석된 단위 피치주기 프레임내의 데이터 갯수가 임의로 81, 81, 82, 83, 82, 84, 82, 81 라고 가정한다. 그러면, 샘플링 주파수 비가 10kHz이면 그림 2(b)에서와 같이, 저장된 고립어의 기본주파수는 123, 123, 122, 120, 122, 119, 122, 123Hz가 된다. 이를 그림 2(c)에서와 같이, 합성하고자 하는 새로운 기본주파수를 지닌 합성음의 피치주기 패턴(130, 135, 132, 130, 128, 126, 125, 118Hz)으로 변경시키고자 하면, 식(10)에서와 같이 저장된 음성에 새로운 갯수열 77, 74, 76, 79, 78, 79, 80, 85 개의 단위 피치 주기 프레임 열로 보간시킴으로써 새로운 피치 주기 열로 변경된 음성파형을 얻는다(그림 2(d)). 그림 5는 음성 파형 "공"을 이러한 피치주기 변경방식을 사용하여 하강음조와 평활음조 및 상승음조를 지닌 합성음으로 변경된 파형에 지속시간과 강약요소를 달리하여 합성시킨 고립어 합성 결과이다.

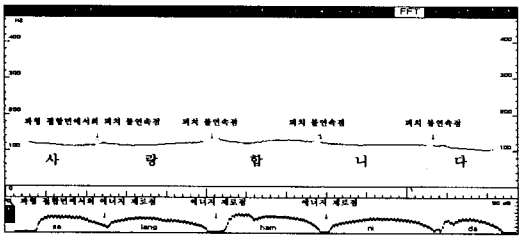
3. 에너지 윤곽선 제어

저장된 고립어 단위의 기본 합성음을 사용하여 시간영역 상에서 합성할 경우에 피치제어 이외에 해결하여야 할 시급한 문제 중의 하나가 음성 파형의 연속적인 에너지 제어이다. 그림 3(a)는 저장된 고립어 단위의 음성 데이터를 연결하였을 경우의 파형을 나타낸다. 이 파형에서는 고립어 마다 각기 다른 고저, 세기, 장

단을 지니고 있어서 청취하면 매우 부자연스럽다. 그림 4(a), (b)는 연속적으로 발성한 음성 "사랑합니다"의 파형으로써 연속음 발성시에는 파형진폭의 흐름이 거의 영으로 떨어지는 일이 없어 연속적으로 이어지나, 고립어 단위에 의한 합성시에는 그림 3(a), (b)의 파형 진폭에서와 같이 음절간 파형 접합점에서 영으로 떨어져 연결되어, 합성음 청취시 음절 사이가 끊겨져 청취되어 매우 부자연스러워진다. 본 논문에서는 이러한 문제를 해결하기 위하여 음성분석에서 추출된 임펄스 진폭 비 매개변수(그림 1에서의 진폭제어용 매개변수 A_{mi})를 이용하여 파형 접합면에서의 에너지 윤곽선의 흐름을 제어하였다(그림 6).



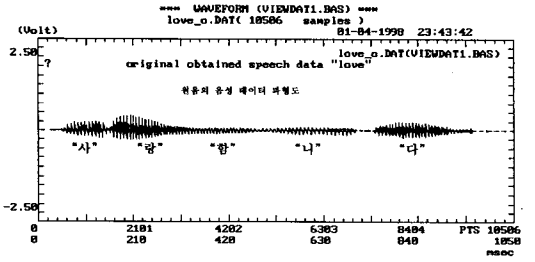
(a)



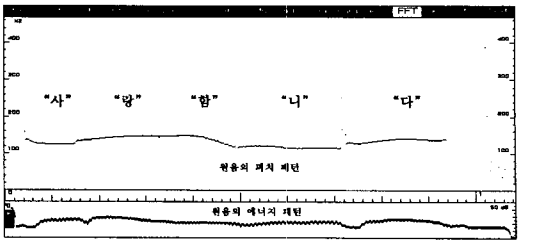
(b)

그림 3. (a) 파형 연결에 의한 음성파형 예 (b) 그림 3(a)의 피치주기 및 에너지 패턴

Fig. 3. (a) The example of concatenated speech waveform (b) The pitch periods and energy pattern of fig. 3(a).



(a)

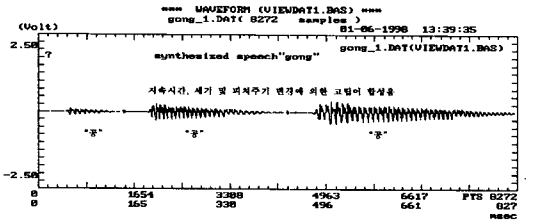


(b)

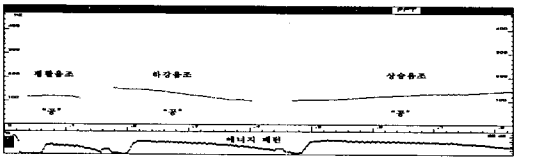
그림 4. (a) 획득된 연속음 데이터 파형 (b) 그림 4(a)의 피치 패턴 및 에너지 패턴

Fig. 4. (a) The obtained continuous speech data waveform (b) The pitch pattern and energy pattern of fig. 4(a).

연속어의 합성 시, 파형 진폭의 연속성은 에너지 윤곽선의 연속성과 일치하게 되므로, 본 논문에서는 음절간 파형 접합점에서의 진폭 포락선의 형태를 연속적으로 제어하였다. 이를 위하여 에너지 윤곽선의 시작점과 끝점을 합성음의 발성 속도 등에 따라 결정한 후, 저장된 임펄스 열의 진폭 비를 조절하여 파형 포락선의 형태를 여러 가지 규칙에 따라 지속시간, 피치 및 에너지의 흐름이 연속적이 되도록 생성하였다(그림 6).



(a)



(b)

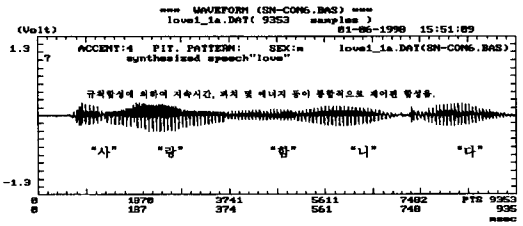
그림 5. (a), (b) 지속시간, 피치 및 세기가 제어된 고립어 합성 예 (a) 지속시간, 세기 및 피치주기가 제어된 고립어 합성 예 (b) 그림 (a)의 피치패턴 및 에너지 패턴

Fig. 5. (a), (b) The example of isolated speech controlled of duration, pitch and stress. (a) The example of isolated speech waveforms controlled duration, stress and pitch periods (b) The pitch pattern and energy pattern of fig. (a).

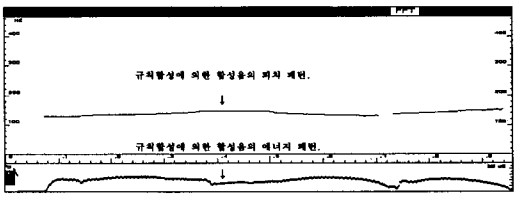
Ⅲ. 실험 결과

1. 운율요소 제어 결과

그림 5(a)는 243ms의 지속시간을 지닌 CVC형태의 고립어 "공"의 음성신호 파형을 150msec, 283msec 및 399msec의 지속시간으로 변경시킨 후 세기와 음조를 각기 다르게 합성시킨 고립어 합성의 한 예를 표시한 것이다. 그림 5(b)는 이의 피치 및 에너지 궤적을 그림으로 나타낸 것이다. 합성시 저장된 고립어의 음성 파형 분석에 의한 운율 제어용 매개변수를 추출하여 단위 음성 DB를 구축한 후, 시간 영역 상에서 이들 운율제어 매개변수들을 이용하여 고립어와 연속음을 합성에 이용하였다. 그림 5(b)에서와 같이 규칙 합성된 음성신호 파형의 피치 궤적도 평활음조, 하강음조, 및 상향음조의 억양을 지닌 피치 주기로 변경되었음을 알 수 있다. 즉, 그림 5(a)는 지속시간과 피치주기 및 세기를 시간 영역에서 제어하여 규칙 합성시킨 결과의 고립어 합성음을 나타낸 것이며, 음질 또한 매우 양호하였다(표 1, 그림 7). 그림 6은 본 논문에서 제안한 시간 영역에서 추출한 규칙 합성용 매개변수를 이용하여 합성시킨 연속음의 합성음 결과이다. 서론에서와 같이 문장 단위의 규칙 합성 시에는 운율요소의 제어가 자연성에 중요한 영향을 미치게 된다. 따라서 자연스런 합성음을 최종적으로 얻기 위 하여는 3개의 운율요소(억양:피치 주기, 강약:에너지, 장단:지속시간)들을 하나로 어우러트려 화자의 감정상태를 자연스럽게 표시하여야 한다. 우선, "사랑합니다."라는 문장을 규칙 합성시킨 그림 6(a)에 대한 합성과정을 기술한다. 이 합성음은 화자가 정상속도로 발성한 경우의 합성음이며, 중심주파수는 120Hz를 기본억양으로 선정하여 합성시킨 결과이다. 첫 번째로 규칙 합성음의 전체지속시간의 결정은 저장된 각 고립어에 대한 전체 지속시간으로부터 발성속도에 따라 합성하고자 하는 각각의 고립어에 대한 단위 피치주기 프레임 갯수에 의하여 결정된다(예를 들면, 정상속도의 발성을 초당 5음절어로 간주 하에 저장된 고립어의 전체 지속시간이 300ms이고, 단위 피치주기 프레임 개수가 30개 이면, 200ms의 지속시간을 지닌 음성 데이터로 변경 시키기 위하여는 단위 피치주기 프레임 갯수를 2/3로 축소시키면 된다). 여기에 피치주기를 변경시키고자 하면 문제는 복잡해 진다. 예를들어 상승음조를 지닌 200msec의 지속시간을 지닌 합성음을 생성시키려면, 상승음조 패턴에 의한 지속시간을 고려하여야 한다. 따라서 본 논문에서는 저장된 매개변수들로부터 수식화 된 데이터 열의 갯수를 정량화하여 변경시키고자 하는 피치주기 패턴에 따라 데이터를 보간시킨 후, 단위 피치주기 프레임 갯수를 조절하여 합성하고자 하는 지속시간을 지닌 음조의 합성음을 연속적으로 생성하여 연속음을 생성하였다. 합성음의 세기(강약)조정은 저장된 음성의 최대 진폭값을 매개변수화하여 세기 정도에 따라 매개변수의 크기를 조절하여 합성하였다. 따라서, 최종적으로 합성하고자 하는 음성의 구문 성격에 따라 장단강약고저 성분 및 음절간 접속구간에서의 파형형태가 결정되면 저장된 매개변수를 호출하여 합성하였다.



(a)



(b)

그림 6. (a),(b) 진폭, 피치 주기 및 에너지 윤곽선 제어 예 (a) 제안된 합성 알고리즘에 의한 연속 합성음 예 (b) 그림 (a)의 피치패턴 및 에너지 패턴

Fig. 6. (a),(b) The example of synthetic speech controlled of duration, pitch period and energy. (a) The example of continuous speech synthesized by presented the algorithm (b) The pitch pattern and energy pattern of fig. (a)

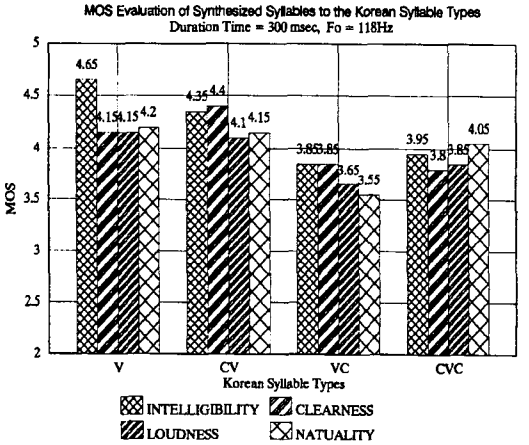
2. 음질평가 결과[10)-14)]

합성음에 대한 음질 평가 방법으로는 합성음에 대한 사전지식이 없는 대학생 각 10명씩 으로 구성된 2개의 그룹(A, B)으로 나누어 음절유형, 지속시간 및 피치변화에 따른 합성음을 4개의 항목(이해도, 명료도, 잡음감, 자연성 등)에 대하여 MOS방법으로 평가하였

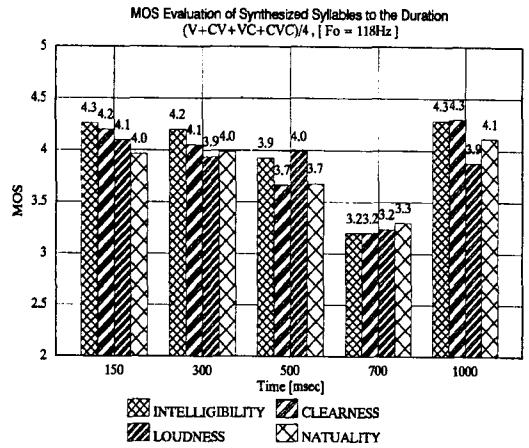
다. 그림 7(a)는 한국어 4가지 음절 유형별로 각각 10개씩 무작위로 추출하여 지속시간을 대략 300ms, 기본 주파수가 118Hz가 되도록 합성시킨 고립어에 대한 평가 결과이다.

표 1. 고립어 합성음 MOS 평가결과
Table 1. The Result of a MOS evaluation to the synthetic isolated words.

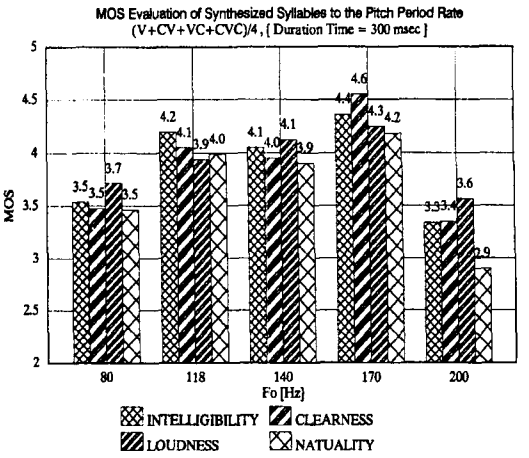
		150msec				300msec				500msec				700msec				1000msec			
		이해도	명료도	잡음감	자연성	이해도	명료도	잡음감	자연성	이해도	명료도	잡음감	자연성	이해도	명료성	잡음감	자연성	이해도	명료도	잡음감	자연성
V형	A	4.3	4.4	3.8	4.0	4.7	4.0	4.0	4.1	4.2	3.4	3.1	3.2	3.8	3.3	4.3	4.1	4.1	4.1	3.8	4.1
	B	4.0	3.7	4.0	3.5	4.6	4.3	4.3	4.3	4.7	4.4	4.7	4.4	3.7	3.1	3.5	3.4	4.7	4.5	4.1	4.3
CV형	A	4.4	4.1	4.1	4.0	4.2	4.2	4.0	3.9	3.0	2.8	2.8	2.8	3.6	3.4	3.7	3.8	4.2	4.2	3.6	4.0
	B	4.2	4.2	4.1	3.7	4.5	4.6	4.2	4.4	4.7	4.4	5.0	4.4	3.8	3.6	3.4	4.0	4.7	4.7	4.2	4.6
VC형	A	3.9	3.9	4.2	3.6	4.2	4.2	3.9	3.8	2.3	2.5	2.9	2.0	2.7	2.4	2.7	2.4	3.8	3.8	3.2	3.5
	B	4.3	4.3	4.2	4.1	3.5	3.5	3.4	3.3	4.7	4.5	5.0	4.6	2.3	2.1	2.2	2.4	4.3	4.3	4.2	4.4
CVC형	A	4.2	4.6	4.1	4.2	3.6	3.6	3.8	4.0	3.1	2.7	3.4	3.3	3.2	3.2	3.4	3.2	4.2	4.0	3.6	3.6
	B	4.8	4.4	4.3	4.6	4.3	4.0	3.9	4.1	4.7	4.6	5.0	4.7	2.5	2.6	2.7	3.1	4.3	4.2	4.3	4.4
계		4.3	4.2	4.1	4.0	4.2	4.1	3.9	4.0	3.9	3.7	4.0	3.7	3.2	3.2	3.2	3.3	4.3	4.3	3.9	4.1



(a)



(b)



(c)

그림 7. (a),(b),(c). 음절유형, 지속시간 및 피치 변화에 대한 규칙 합성음 음질 평가 결과
(a) 한국어 음절 유형별 규칙합성음 음질 평가 결과 (b) 한국어 지속시간 변화에 대한 규칙 합성음 음질 평가 결과 (c) 한국어 피치변화에 대한 규칙합성음 음질 평가 결과

Fig. 7. (a),(b),(c). The result of a MOS evaluation to the Korean syllable types, durations and pitch variations.
(a) The result of a MOS evaluation to the Korean syllable types (b) The result of a MOS evaluation to the durations. (c) The result of a MOS evaluation to pitch variations.

그림 7(b)는 한국어 4가지 음절 유형별로 각각 10개씩 총 40개의 고립어를 5가지의 지속시간을 지닌 고립어로 합성하여 청취도 실험을 행한 결과이다. 그림 7(c)는 300ms정도의 지속시간을 지닌 고립어를 80Hz에서 200Hz정도의 중심주파수를 지닌 5가지 유형의 피치패턴으로 변경시켜 합성시킨 합성음에 대한 평가 결과이다.

Ⅳ. 결 론

본 논문은 시간 영역에서의 운율요소 제어에 관한 연구로서 저장된 음성 데이터로부터 매개변수를 추출하여 제어시키는 방식을 제안하였다. 그 가능성에 대한 타당성 검토로 3장에서 실험 결과를 제시하였다. 합성시 매개변수를 수식화하여 운율요소(장단강약고저)를 제어함으로써 통합적인 운율요소의 제어가 가능하여져 자연성이 향상되었다. 합성을 위한 초기 단계로서 파형분석시, 피치 주기 추정이 잘못되어진 매개변

수가 음성 DB에 수록되면, 합성시에는 음질에 커다란 영향을 미친다. 따라서 피치 주기를 정확히 추출하여 음성 DB를 구축하여 합성음을 생성하여야 한다. 실험 결과, 4가지 유형의 한국어 고립어 합성 시에는 장음과 단음 같은 합성음은 지속시간에 거의 제한이 없는 합성음이 생성되었다. 저장된 기본 단위합성음으로부터 시간영역에서 억양요소 등과 같은 운율요소를 제어할 경우 가장 큰 문제점은 외부함수를 가하여 인위적으로 파형을 변화시킬 경우 고유의 음질이 훼손되는데 있다. 따라서 저장된 원음의 훼손을 방지하고 운율요소의 효율적인 제어를 위하여 피치 주기 변경시 라그랑쥐 3차 보간법을 이용하여 제어하는 방법을 제안하였다. 실험 결과, 라그랑쥐 보간법을 사용함으로써 생성시키고자하는 임의의 피치 주기 궤적에 따라서 상향억양, 하향억양등 원하는 억양을 지닌 규칙합성음을 자유로이 생성 가능 하였다. 음질 또한 매우 자연스럽고 양질의 음질을 지닌 합성음으로 생성 가능하였다. 저장된 음성 파형으로부터 시간 영역에서 음조를 높여 규칙 합성시킬 경우, 기본주파수 200Hz 이상의 고음조로도 변조가 가능하였으며, 여성 음에 대한 피치 주기 검출 후 규칙 합성시킨 경우도 남성 음의 경우와 동일한 결과를 얻었다. 다음절어 합성 시 음절간 접속구간에서의 파형 윤곽선 제어 및 변이음 처리가 개선되었다.

다음절어 및 문장단위의 합성을 위하여는 한국어에 대한 언어학적인 연구 즉, 한국어 구문 및 문장 단위에서의 음절별 지속시간, 진폭, 피치 패턴 등과 같은 수많은 연구가 앞으로 시급히 이루어져야 보다 자연스런 합성음을 지닌 한국어 TTS시스템의 구현이 가능해질 것이다.

참 고 문 헌

[1] 이현복, "현대 한국어의 악센트," 서울대학교 문리대학보 19권 합병호(통권28호), 1973

[2] 성철재, "표준 한국어 악센트의 실험 음성적 연구 - 청취 테스트 및 음향분석," 서울대학교 대학원 석사학위 논문, 1991

[3] 이현복, 한국어의 표준발음, 교육과학사, 1989

[4] Jonathan Allen, M. Sharon Hunnicutt and Dennis Klatt, From Text to Speech : The MITalk system, Cambridge Univ. Press, 1987.

[5] Shuzo Saito, Fundamentals of Speech Signal Processing, Academic Press, 1981.

[6] G. Rigoll, "The DECtalk system for German : A study of the modification of a text-to-speech converter for a foreign language," IEEE Proc. ICASSP '87, 1987.

[7] T. Dutoit, H. Leich, "Improving the TD-PSOLA Text-To-Speech Synthesizer with a Specially Designed MBE re-Synthesis of the Segments Database", ICASSP 92, vol. 1, pp. 343-346.

[8] T. Dutoit, "High Quality Text-To-Speech Synthesis: a Comparison of Four Candidate Algorithms", Proc. ICCASSP 94, vol. 1, pp. 565-568.

[9] 中律 良平, "音聲認識·合成技術の製品化および"市場動向," SP89-103, 1989

[10] 한국방송공사, 표준한국어 발음 대사전, 어문각, 1993

[11] Nobuhiko Kitawaki, Hiromi Nagabuchi, "Quality Assessment of Speech Coding and Speech Synthesis System," IEEE Comm., 1988. vol. 26. no. 10.

[12] Toshiro Watanabe, "規則合成音の自然性評價法の檢討", 電子情報通信學會論文誌, A vol. J74-A no. 4, 1991

[13] 조철우, 김경태, 이용주, "무의미단어에 의한 규칙 합성음의 평가 및 진단법에 관하여," 음성통신 및 신호처리 워크샵 논문집, 1993. 8

[14] 김정환, 강성훈, "음성품질 주관법의 표준화에 관한 고찰," 전자통신 동향분석, 1990. 7

저 자 소 개

姜 贊 熙(終身會員)

1980년 2월 경희대학교 전자공학과 졸업(학사). 1982년 2월 경희대학교 대학원 전자공학과 졸업(공학석사), 1983년 7월 ~ 1985년 7월 해군사관학교 교수부 전자과 전임강사. 1985년 8월 ~ 1986년 8월 삼성 통신 연구소근무(연구원). 1986년 9월 ~ 1994년 8월 경희대학교 대학원 전자공학과 졸업(공학박사). 1989년 3월 ~ 현재 상지대학교 병설 전문대학 전자과 부교수. 주관심 분야는 한국어 TTS 시스템내 음성합성 알고리즘 개발 및 구문분석기 설계, ARS 시스템 구현, 음성인식 및 디지털 음성 신호 처리 등임