論文 90-27-10-18

반음절단위를 이용한 한국어 음성합성에 관한 연구

(A Study on the Korean Text-to-Speech Using Demisyllable Units)

尹 起 善*, 朴 成 漢**

(Gi Sun Yun and Sung Han Park)

要 約

본 논문에서는 합성단위를 반음절로 하여 적은 데이터 베이스를 차지하면서도, 합성음의 자연스러움을 향상 시키기 위한 한국어 규칙 합성법을 제시한다.

반음절 음성신호를 분석하기 위해 12차 선형 예측법을 사용하며, 합성음의 자연성과 명료성을 위해 음절간 접속 규칙, 모음부의 연결규칙을 개발한다. 또한 신경망 모델을 이용한 음운 변동 규칙과 운율규칙을 적용한다.

Abstract

This paper present a rule-based speech synthesis method for improving the naturalness of synthetic speech and using the small data base based on demisyllable units.

A 12-pole Linear Prediction Coding method is used to analyses demisyllable speech signals. A syllable and vowel concatenation rule is developed to improve the naturalness and intelligibility of the synthetic speech. In addiion, phonological structure transform rule using neural net and prosody rules are applied to the synthetic speech.

I. 서 론

음성 합성은 1971년 벨 연구소에서 처음으로 디지탈 음성 합성기를 발표한 이래 인간의 가장 기본적이고 편리한 통신 수단인 음성을 이용하여 컴퓨터와 인간의 정보 전달을 수행하기 위해 여러 분야에서 꾸준히 연구되어 오고 있다. 지금까지는 기계로부터 인간으로의 통신수단이 주로 시각을 통하여 정보를 전달하는 방법이 사용되었으나, 정보의 형태가 다양해짐에 따라 시각적인 방법만으로는 불편하게 되었다. 그 대안으로서 무제한 어휘의 음성합성이 연구되고 있다. 또한 반도체 기술의 발달에 힘입어 기억용량에의 제약이 점차 적어짐에 따라, 전자 사서함, 전화번호 안내시스템, 일기예보 teaching machine, 음성에의한 경고 시스템등 그 응용 분야가 광범위하다. 이러한 합성의 단위로는 음소, diphone, 반음절, 음절 등이 있다.[1-3] 음소단위는 음소 각각의 지속시간, 기본 주파수, 강세등의 특성변화를 쉽게 적용시킬 수 있으나, 음소간의 조음 결합처리가 어려운 단점이 있고, diphone은 한음소의 중간과 다음 음소의 중반까지의 음성신호로 이루어진 것으로 인접 음소간의 조음 결합이 원활한 장점이 있지만, 모음간에 위치

*正會員, 現代電子通信開發 1 部
(HEI Telecom. Systems Division R&D Dept. 1)
**正會員, 漢陽大學校 電子計算學科
(Dept. of Computer Science and Engineering, Hanyang Univ.)
接受日字 : 1990年 4月 6日

하는 자음을 중심으로 하는 모음사이의 조음 결합이나, 강세등의 음운적 환경에 따른 특성변경이 용이하지 못한 단점이 있다. 반음절은 음절을 반으로 나눈것으로 음절단위 합성에서와 같은 자연스러움을 얻을 수 있으며, data base의 양도음절에 비해 매우 적고, 제어도 쉽다.

음성의 규칙합성은 순수한 규칙만으로 음성을 합성하는 방식과 vocoder를 이용하는 방법이 있다. 전자의 경우는 순수한 규칙으로 합성하므로 데이터 베이스의 양이 적으나 국어에 대한 과학적인 연구와 그리고 많은 분석과 교정을 필요로 하며 시스템도 복잡해지므로 이를 실제적으로 음성합성에 적용하기에는 많은 어려움이 따른다. 후자는 데이터 베이스의 양은 다소 많으나 비교적 고품질의 합성음이 얻어지고 시스템도 비교적 실제적이다. 그러나 우리나라에서는 아직 국어에 대한 과학적 연구가 부족하여, 한국어를 유창하고, 자연스럽게 할성할 수 있는 규칙을 정립할 수 없는 상태이다. 따라서 본 논문은 한국어 text에 대하여 음성으로 합성하는 한국어 text-to-speech에 대한 연구로서, 명료성을 해치지 않는 범위내에서 사람의 음성과 비슷한 자연스러움을 가지는 규칙을 개발한다. 기본적인 합성방법으로 분석 합성 방법인 선형 예측(linear prediction) 분석 합성법을 이용하고, 단일화자의 음성 신호를 반음절 단위로 분할(segmentation)분석하여, 선형 예측계수를 구하고, 이 계수를 이용하여 음성을 합성한다.

규칙합성을 수행하는 경우 철자(음소)에 대해서 단어나 혹은 같은 문장속에서도 쓰이는 역할에 따라 다르게 발음되는 현상이 존재한다. 이러한 문제점을 해결하기 위하여 음성학적 발음 특징에 관한 연구를 토대로 발음규칙을 설정하여 발음규칙에 의해, 입력된 문장들을 보다 정확한 발음표기로 변환시키는 시스템들이 필요하다.

이를 위하여 본 논문에서는 신경망 모델을 이용하여 음운 변동을 수행하며, 인공 신경망 모델은 음성과 영상 인식분야에서 인간이 갖고 있는 자연정보처리(natural information processing) 기능과 판단능력을 컴퓨터에 부여하려는 노력으로 고안된 모델로서, 기존의 신호처리 방식에 의해서는 처리하기 어려운 자연 정보처리 기능과 불완전한 정보의 확률적인 처리, 그리고 고도의 병렬 분산처리 효과등으로 인해 음성정보의 실시간 처리가 가능하다.

최근의 새로운 NET와 학습 알고리듬의 개발, VLSI로의 실현, 음성이나 영상인식에서 고도의 수행 능력을 얻기 위해서는, 고도의 병렬성을 갖는 모델이 연구되고 있다. 그 예로 Johns hopking 대학의 sejnowski와 princeton 대학의 rosenberg가 개발한 NET talk는 영어 문장의 발음기호 생성 시스템으로 신경망 응용의 대표적인 시스템이다. 따라서 본 논문에서는 NETtalk[10,11,12]를 응용해서 한국어 문장을 한국어 음운 규칙에 따라 발음기호로 출력해 주는 인공 신경망 모델을 이용하여 음운변동을 수행한다. 발음기호나 국어의 코딩 방법은 compact code 방법[12] 을 이용하였으며 89% 정도의 성취도를 나타내었다.

Ⅱ. 반음절 규칙 합성 시스템

Text-to-speech 시스템은 임의의 입력문장에 대해 그에 대응하는 음성신호로 출력하는 것이므로, 이를 수행하기 위하여는 몇가지 전처리 단계가 필요하며, 적은 데이터 베이스를 차지하고, 합성음을 쉽게 제어할 수 있는 vocal track model을 이용한 방법인 LPC 또는 FORMANT 합성 방법을 많이 이용하고 있다. 이러한 text-to-speech 시스템을 구현하기 위하여 본 논문에서는 그림1과 같은 절차를 따라서 수행한다.

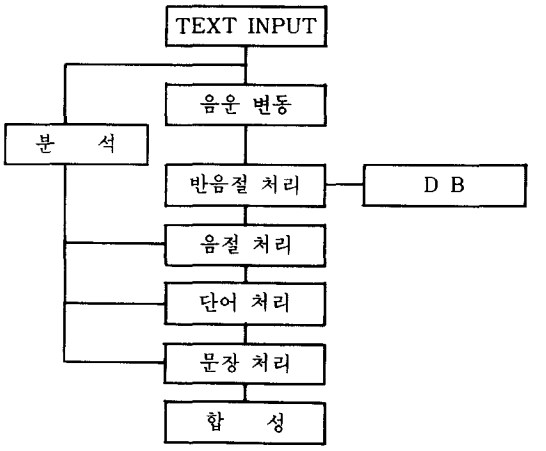


그림 1. text-to-speech의 구현을 위한 흐름도
Fig. 1. The flowchart of text-to-speech processing

그림 1에서와 같이 임의의 tevt가 입력되면, /, /, /. /, /? /, /! /등과 숫자등을 음운 변동 시스템에서 필요한 코드로 변환해주며, 음운 변동 처리부에서는 한국어 문장을 소리나는대로 바꾸어 주며, 평문, 의문문, 느낌문, 감탄문에 대한 문의 종류와 문장내에서의 구의 갯수등 음율처리에 필요한 정보를 추출한다. 또한 반음절 처리 단계에서는 음절의 형태가 "각," "악," "아," "가"와 같은 초성+중성+종성, 중성+종

성, 종성, 초성+중성 형을 구분하고, 초성+중성+종성의 형태일 경우 초성+중성과 중성+종성의 반음절 파라미터를 선형 보간과 스므딩(smoothing) 방법을 이용해서 연결하고, 중모음, 복모음의 모음절 조절을 위해서 '단모음+반모음'인 경우는 기본주파수 형태가 하강하는 패턴 '반모음+단모음'인 경우는 상승하는 패턴을 갖도록 데이터 베이스를 제어한다.[7,8] 음절처리 단계에서는 음절 연결규칙에 따라 이전 음절의 종성 부분과 현재 음절의 초성부분을 연결한다. 또한 단어처리 단계에서는 악센트 규칙에 따라 악센트를 부여하며, 음절수에 따라 frame의 갯수를 조절하여 합성음을 자연스럽게 한다. 문장처리 단계에서는 의문문, 평서문, 감탄문, 명령문에 따라 문장의 운율과 어조등을 처리한후, 디지탈 합성기를 통해 음성을 합성한다.

Ⅲ. 신경망을 이용한 음운 변동

한국어에 있어서 text의 문장과 text를 발음한 경우의 문장과는 음소들이 차이를 나타내고 있다. 즉 음절이 모여 낱말로써 발음될때, 음절과 음절이 잇달아 소리나거나 단어와 단어가 잇달아 소리나면 서로 영향을 주고 받아서 여러가지 다른 소리로 발음된다. 이러한 현상을 고려하여 text로 표기된 것을 발음나는 대로 바꾸어주는 한국어 음운 변동 처리시스템이 필요하다.[1] 본 논문에서는 이의 구현을 위해 자연 정보처리 기능과 불완전한 정보의 확율적인 처리, 그리고 고도의 병렬 분산처리 효과등으로 인해 음성정보의 실시간 처리가 가능한 신경망 모델을 이용하여 음운 변동을 처리한다.[10,11,12]

1. 신경망 모델

인공 신경망 모델은 생물학적인 형태로 배열되어 있고, 병렬로 동작하는 많은 비선형 노드들로 구성한다.

최근의 새로운 NET와 학습 알고리듬의 개발, VLSI로의 실현, 음성이나 영상인식에서 고도의 수행 능력을 얻기 위해서는, 고도의 병렬성을 갖는 모델이 연구되고 있다. 그 예로 Johns hopkins 대학의 sejnowski와 princeton 대학의 rosenberg가 개발한 NETtalk는 영어 문장의 발음기호 생성 시스템으로 신경망 응용의 대표적인 시스템이다. 따라서 본 논문에서는 NETtalk[10]를 응용해서 한국어 문장을 한국어 음운 규칙에 따라 발음기호로 출력해 주는 인공 신경망 모델을 이용하여 음운변동을 수행한다. 신경망의 기본구조는 그림 2 와 같다.

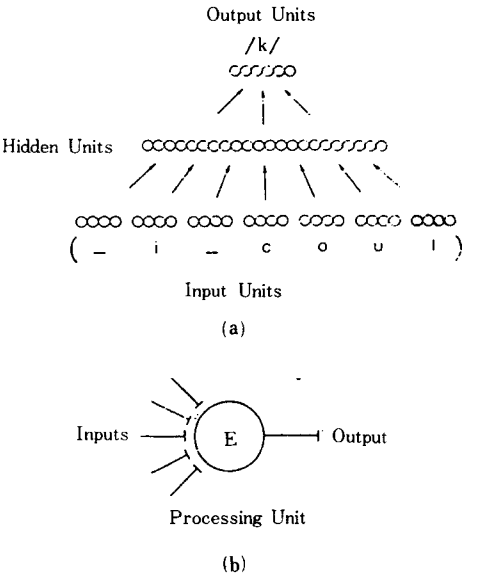


그림 2. (a) NETTalk의 기본 구조
(b) 처리 단위
Fig. 2. (a) Basic struct of NET Talk,
(b) Processing Units.

그림 2(a)에서 보는 바와 같이 입력 layer의 가운데 window 'c'가 출력 layer의 출력 'k'로 나타난다. 입력 layer에서 'c'의 발음은 좌우 3음소의 영향을 받는다고 간주하였기 때문에, 입력에서 ~7개의 입력 window가 필요하다.[10,11]

그림 2(b)에서 보는 바와같이, 신경망 구조는 연속된 입력에 대해 입력의 가중치 합을 비선형적으로 변환하는 처리 unit로 구성되어 있다. 한 unit와 또 다른 unit 사이의 연결 강도(weight)들은 양, 음의 실수치를 갖고 그 값에 따라 자극을 억제하기도 하고, 자극이 주어지기도 한다.

그림 2(b)에서 임의 node의 출력은 그 노드로 향한 모든 입력의 합, E_i^t가 되며, 그 합에 그림 3의 전달함수가 적용, 그 node의 실제 출력, p^t가 구해지며, 이출력은 상위노드의 입력이 된다.

즉

$$P^t = P(E^t) = \frac{1}{1+e^{-E^t}}$$

또한 이 신경망의 수렴여부는 각 노드간의 연결강도에 따라 좌우되며 주어진 망입력에 대응하는 망출력을 내도록 연결 강도를 조절하는 과정이 필요한데, 이러한 과정을 학습이라 하며 신경망의 설계와 함께 현재 연구의 주된 테마이다.[10,11]

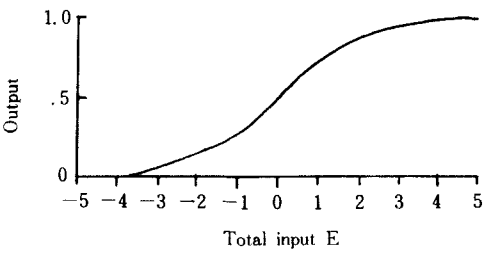


그림 3. 문턱치
Fig. 3. Threshould logic.

2. 신경망의 학습규칙

Back-propagation 학습규칙은 원하는 출력과 신경망의 망출력 사이의 오차를 최소화 하기 위하여 고안된, 반복적인 경사도(iterative gradient) 알고리듬이다. 이는 연속적이고, 미분가능한 비선형성이 필요하며 그 비선형성은 sigmoid 함수가 사용되고 있다. 즉

$$f(E)=\frac{1}{1+e^{-(E-\theta)}}$$

여기서 E=노드의 net output

θ=문턱치

그림3는 이러한 비선형 함수를 나타내고 있다. 이의 학습과정을 살펴보면 아래와 같다.[11,12]

1. 연결강도와 문턱치를 작은 난수로 초기화한다.
2. 연속적인 입력과 그에 대응하는 출력을 신경망의 입출력층에 부여한다.
3. 식에 있는 sigmoid 비 선형 함수를 이용해 각 노드의 망 출력을 구한다.
4. 출력 층에서부터 시작하여 hidden layer, hidden layer에서 입력층으로 순환 알고리듬을 적용연결 강도를 조절한다.

   $W^{ij}(t+1)=W^{ij}(t)+\delta^j\times^i$

   윗식은 연결강도의 조절 방법을 나타내는 식으로 $W^{ij}(t)$는 시간 t일때 node i에서 node j로의 연결강도 이고, x는 node i에서의 망출력이고, δ^j는 node j에서의 오차이다.

   만일 node j가 출력 노드이면

   $\delta^j=y^j(1-y^j)(d^j-y^j)$

   여기서 d^j는 노드 j에서의 원하는 출력이고, y^j는 노드 j에서의 망 출력이다.

   만일 노드 j가 내부의 hidden 층이면

   $\delta^j=x^j(1-x^j)\ kW^{jk}$,

   여기서 k는 노드 j의 상위층의 전 노드이다.
5. 원하는 출력과 망 출력 사이의 오차가 충분히 작지 않으면 2단계로 간다.

3. 음운 변동 시스템

한국어 문장을 발음기호로 변환 시키기위한 신경망 모델은 그림4와 같다.

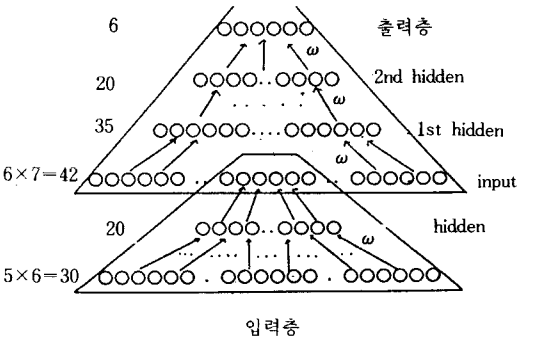


그림 4. 임의의 한국어 문장을 발음기호로 변환하는 NET
Fig. 4. The NET to convert given korean sentences.

그림4의 윗 부분에서 입력노드는 7개의 입력창으로 되어 있으며 각각의 창에 한 음소씩 할당한다. 이는 한글자소의 음운변화 영향권을 최대 3자소로 간주 하였기 때문이다. 따라서 가운데 window의 음소가 좌우 3개의 음소의 영향을 받아 출력 layer로 발음기호가 출력된다.

한개의 입력창은 6개의 노드로 구성되어 있어 입력 노드가 총 42개로 구성되어 있으며, hidden layer는 한개의 창으로 이루어져 있고, 35개의 노드를 가진 첫째 layer와 20개의 노드를 가진 두번째 layer로 구성되는 두개의 hidden layer를 두어 폭넓은 규칙들을 내재시키고, 예외 규칙들은 특수화 시켜 스스로 학습하도록 하였다. 출력은 한개의 창으로 되어 있으며 6개의 노드로 구성되어 있다. 그러나 이러한 net에서 저장할 수 있는 정보가 한정되어 있고, 발음규칙을 학습하는 데에 복잡성이 높으므로 예외 규칙들이 많은 종성규칙을 새로운 망에 부여하여 처리하도록 하므로써 성취도를 보다 향상시킬 수 있었다.

그림4의 아래 부분은 이러한 종성 규칙을 처리하기 위해 사용된 신경망을 나타낸다. 이 망은 30(6×5)개의 입력 unit와 20개의 hidden unit, 그리고 6개의 출력 unit로 구성되고 이망의 출력은 곧바로 상위 네트의 입력 unit로 입력이 된다. 또한 모든 입력 노드에서 입력의 형태는 종성 규칙에 서로 밀접한 음소들에 순차적으로 코드를 부여해서 시스템의 성능을 향상시키도록 한다. 총 음소가 33개 이므로 $\log_2 3^3=6$이므로, 6비트이면 모든 음소를 코딩할 수 있다.

Ⅳ. 데이타 베이스의 작성

무제한 음성합성에서 합성단위의 선택은 매우 중요하다. 이러한 합성의 단위로는 문장, 단어, 음절 반음절, 음소등이 있다. 문장이나 단어의 경우, 음성의 정보가 거의 들어 있으므로 합성음의 질이 좋고 자연스러운 음을 가지나, 무제한 음성합성에 적용하기에는 그 차지하는 데이터 베이스의 양이 너무 많으므로 text-to-speech에 적용하기에는 상당한 무리가 따른다.[2]

또한 합성단위는 음절에서 반음절, 음소로 갈수록 차지하는 데이터 베이스의 양은 적고, 합성음의 음질은 다소 나빠진다. 따라서 데이터 베이스의 양과 합성음의 음질을 고려해서 적절한 합성단위의 선택이 요구된다.

본 논문에서는 합성음의 질은 다소 떨어지나, 데이타 베이스의 양이 적은 반음절을 합성단위로 한다.

그림 5는 LPC방법에 의한 반음절단위 데이터 베이스의 작성을 위한 전체 흐름도이다.

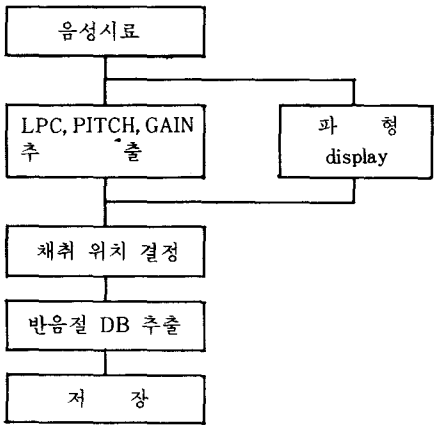


그림 5. 데이터 베이스 작성을 위한 흐름도
Fig. 5. Flowchart to make speech database.

주어진 음성 시료에서 pitch와 gain, LPC 계수를 구하고, 파형을 보고 채취할 데이터의 위치를 구한다. CV형인 경우 음성시료의 시작 부분부터 transition 이후 부분 까지이고, VC형인 경우 transition 이 일어나기 이전 부분부터 음성 시료의 끝 부분까지 채취한다. 그리고 V형인 경우 정상 상태의 일부를 채취하여 디스크에 저장한다.

Ⅴ. 음절연결규칙

두음절간의 음소들의 결합의 제약성과 상호 작용으로 인해 합성음의 음질에 큰 영향을 미치므로 자연스러운 연결을 위하여 연결규칙이 필요하다. 따라서 본 논문에서는 이러한 연결 규칙을 구현하기 위하여 선형보간, SMOOTHING, DIRECT 등의 연결 방법을 사용하여 음절 연결 규칙을 수행하며 음절사이의 연결을 자연스럽게 한다. 우리말의 음절의 됨됨이를 보면 아래와 같다.[7,8]

Ci V Cf (Ci=초성자음, V=음절핵, Cf=종성자음)

[Ci]의 위치에는 19개의 자음이 다올 수 있으나, 음절의 위치에 따라서는 약간의 제약이 있으며, [V]의 위치에는 10개의 단모음과 12개의 중모음이 올 수 있다. [Cf]의 위치에는 7자음만이 올 수 있다.

그러므로 음절의 구성은

CVC =2926
CV 19*22=418
VC 22*7 =154
V =22

본 논문의 합성 단위가 반음절이므로 총(418+154+22=594)개의 데이타 베이스만 있으면, 무제한 한국어 규칙 합성을 수행한다. 그러나 음소 결합의 제약이 있으므로 실제적으로는 이보다 더 줄어들게 된다.

1. 정상 상태 부분의 연결

합성단위가 반음절이므로 음절을 이루기 위해서는 두개의 반음절의 연결이 요구된다. 이와같이 반음절을 연결할 경우에 있어서 연결규칙이 필요하다.[7]

가) 다음 음절이 space일 경우→ 개방연접은 반음절 연결시 정상상태 부분을 길게 유지하므로서 구현한다.

나) 장모음과 중복모음의 경우→ 중복모음과 장모음, 이러한 경우 중복모음은 단순한 긴소리가 아니고 그사이에 tension의 하락점(음절의 경계)을 두어 두음절로 발음하며, 장모음은 tension의 하락점이 없이 길게 발음하여 구현하며 tension의 하락점은 gain과 pitch로 조절한다.

다) 중모음인 경우→ '반모음+단모음' 일때는 모음점이 뒤에 있기 때문에 상승점 이중모음이고, '단모음+반모음'은 모음점이 앞에 있기 때문에 하강적 이중모음이 되어 두 반음절 연결시 pitch의 상승 하강 조절이 필요하다. 예로 그림 6에서와 같이 "야"는 반모음+단모음의 형태로 단모음인 "a"에 조음점이 있다.

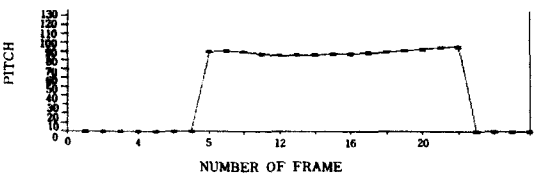


그림 6. 중모음 '야'에 대한 피치' 패턴
Fig. 6. Pitch pattern of '야'.

2. 음절 연결 규칙의 구현

음절 연결규칙을 위해 연결이 가능한 음소가 있는 reference data를 분석하여 아래와 같이 구현하였다.

1) 종성 ···< /p'/,/t'/,/k'/
2) /p/,/t/,/k/ ···< /p h/,/t h/,/k h/
3) 〃 ···< /c'/
4) 〃 ···< s'
5) 〃 ···< empty
6) empty ···< 유성자음
7) empty ···< /p/,/t/,/k/
8) 〃 ···< /p'/,/k'/,/t'/
9) 〃 ···< /p h/,/k h/,/t h/
10) 〃 ···< /c/,/c'/ / /
11) 〃 ···< 유성자음
12) /l/,/m/,/n/,/ / ···< /p/,/t/,/k/
13) 〃 ···< /p'/,/t'/,/k'/
14) 〃 ···< /p h/,/k h/,/t h/
15) 〃 ···< /s/,/s'/
16) 〃 ···< /c/,/c h/,/c'/

1),2)는 1frame 정도의 휴지를 삽입시켜 성도를 폐쇄하는 효과를 내고, direct 연결한다.

3),4),13),14)은 gain의 스므딩후에 direct 연결한다.

5)는 3frame 정도의 선형 보간을 하며 음절의 경계를 두기 위해 점강, 점약음의 형태가 되게한다. 또한 pitch를 스므딩해 갑작스러운 vocal tract의 변화를 없앤다.

6)는 3frame 정도의 선형 보간을 하며 음절의 경계를 두기 위해 점강, 점약음의 형태가 되게한다. 또한 pitch를 스므딩해 갑작스러운 vocal tract의 변화를 없애고, 유성자음에 악센트를 없애기 위해 피치의 조절을 하며, 유성자음의 길이를 선형보간에 의해 늘인다.

7),12)는 앞음절이 점약음의 형태가 되게하며, 뒤음절이 율성음화 되어 성도의 폐쇄가 없으므로 데이터 베이스에서 gain이 갑자기 커지는 부분까지 제거한 후 연결된다.

8),9)는 앞음절이 점약음 형태가 되게하며, 1frame 정도의 휴지로 성도를 폐쇄시킨다.

10)은 앞음절이 점약음이 되고, 앞음절의 유성으로 인해 뒤음절의 초성이 잘 들리지 않으므로 1frame 정도의 noise를 삽입한 후, 뒤음절의 자음의 길이를 늘여 연결한다.

11)2프레임 정도의 선형보간을 행하고, gain이 점약, 점강의 형태가 되도록 한다.

15)는 앞음절이 유성음이므로 뒤음절의 /S/음이 잘 들리지 않으므로 약간의 noise와 /s/음의 길이를 interpolation에 윗해 늘인다.

16)점약음 형태가 되게 하며, 유성자음의 길이를 늘인다.

그림 7은 이와 같은 음절 연결 규칙에 의해 연결된 "산에 가십니까?"의 피치 패턴을 나타내고 있다.

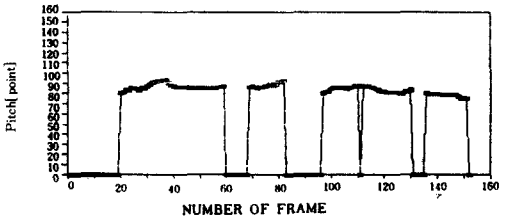


그림 7. "산에 가십니까?"에 대한 피치 패턴
Fig. 7. Pitch pattern of "산에 가십니까?"

Ⅵ. 운율 규칙

음성의 음율은 기능면에서 듣는 사람으로 하여금 메세지에 대한 구조적인 정보를 제공한다. 이러한 음율은 음성의 템포, 리듬, 멜로디를 형성하는 음향학적인 요소는 intensity, stress, pitch contour등에 의해 표시된다.

억양은 강세와 상조 체계에 관련된 구실외에도 모든 언어에서 구절과 문을 구별하고 경계를 표시하는 기능을 가지며, 음절보다 큰 단어나 문장단위에 그 영향이 확장되는 고저의 차이로 표현되며 문장 전체의 의미 파악에 중요한 역할을 한다.

한 단어 안에서 음절의 상대적 현저도(prominence)를 강세라고 한다. 영어를 포함한 많은 언어에서 단어속의 어떤 음절은 다른 음절보다 상대적으로 더 우세한 현저도를 갖는다. 국어에서 단어의 강세는, 일반적으로 받침이 있는 음절에 오며, 받침이 없는 단어의 경우 대개 첫음절에 강세가 있다고 보고되어있다.[15]

또한, 음량은 분절음의 상대적 길이를 말한다. 어떤 분절음은 실제적으로 그 음을 산출하기 위하여 다른 것보다 더 길게 발음한다. 전형적으로 긴장모

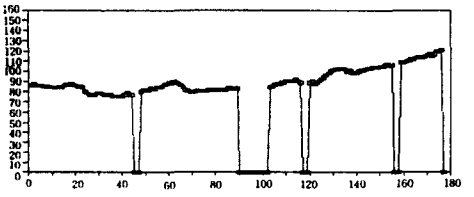


그림 8. "우리는 서로 도와야 한다"의 피치 패턴
Fig. 8. Pitch pattern of "우리는 서로 도와야 한다."

음은 이완모음보다 더 길게 발음하며 저모음, 고모음보다 더 길게 발음하고, 이중모음이 단모음보다 더 길어진다. 그러나 대부분의 음성은 그들 본래의 음장에 관계없이 상대적으로 더 길거나 더 짧은 형식으로 산출될 수 있고, 말소리의 상대적 음장 혹은 음량은 대별 언어들에서 단어의 의미를 변별하는데 자주 이용된다.

또한 국어는 다른언어(특히 영어)와는 달리 음절의 응량, 즉 상대적 길이를 잘 나타내야 발화의 흐름이 잘 되어지는 음절, 시간 리듬(syllable-timed rhythm)의 언어에 속한다. 영어와 같은 언어는 발화의 흐름이 일정한 강세. 시간리듬(stress-timed rhyth을 가지고 있어서 강세리듬을 바로 나타내어야 발화의 흐름이 잘되어 유창한 말로 인식된다.[7,8]

소리의 높낮이가 통사적 대립에 이용되는것을 월가락(억양)이라 한다. 국어의 구말 고저는 세가지가 있어서, 구절의 끄트머리는 상승, 수평, 하강의 어느 한편으로 발음되며, 이러한 음저는 말의 뜻을 알아듣는데 중요한 구실을 할 수 있을뿐만 아니라 때로는 말의 뜻을 분화하는 구실을 하는 일도 있다. 예로 [해↑]는 [할까?]와 같은 의문이며 [해↓]는[해라!]와 같은 명령이다. 또 하강의 경우에도 그 가락을 빨리 낮추면 서술이 된다.

수평 고저는 대체로 말이 이어나갈 경우에 쓰이며, 하강 고저는 서술과 명령에 쓰이며, 그리고 상승 고저는 의문을 표시하게 된다. 국어의 어중에 있어서의 고저는 대체로 평탄함이 그 특색이며, 그 고저의 차가 생겨날때는 감정의 고저가 있음을 의미하는 것이 보통이다.

따라서 본 논문에서는 schemetic 알고리듬을[9,14] 이용하여 이를 실현하며 음절, 단어로 발음된 경우와 문장으로 발음된 경우의 reference data를 분석하여 본 결과 음절, 단어에 비해 문장으로 발음된 data가 정상상태 부분이나 음절의 길이가 비교적 짧은 것으로 나타났다. 이는 음절수가 많은 단어나 문장을 발음할 경우 다르게 발음하는 경향이 있기 때문이다. 따라서 단어에서의 음절수를 계산하여 음절수가 많을수록 상대적으로 음절의 duration을 줄이는 방향으로 하며 이는 모음의 정상상태 부분의 frame수를 줄이므로써 실현할 수 있다. 그림 8는 이러한 강세, 음량, 억양의 규칙이 적용된 "우리는 서로 도와야 한다"의 피치 패턴을 나타내고 있다.

Ⅶ. 실험결과

본 연구에서는 반음절을 합성 단위로 하여 한국어 문장을 분석합성법에 의해 합성하며, 자연스러운 합성음의 출력을 위해 신경망 모델을 응용한 음운변동 시스템, 음절 연결규칙, 반음절에서의 모음간 접속 규칙및 음율조절 규칙을 적용하였다.

합성단위에 있어서 음절로 하였을 경우 그 차지하는 데이터 베이스의 양이 너무 많으며, 데이터 베이스를 제어하기 어렵다. 음소단위의 경우 자음과 모음의 데이터 베이스만 만들면 충분하나, 아직 국어의 음소에 대한 연구가 충분치 않아, 무제한 음성합성을 행하기에는 현실적으로 많은 어려움이 따른다. 그러나 반음절 단위를 사용시 500개 정도의 데이터 베이스만으로 모든 음을 합성할 수 있으며, 음절과 유사한 자연스러움 음을 얻을 수 있었으며, 음절에 비해 데이터 베이스를 제어하기 쉬운 장점이 있다.

신경망 모델을 이용한 음운 변동 시스템은 현재 한국어에서 자주 쓰이는 단어 1080개를 채취하여 그중 500개 정도는 신경망을 교육시키는데에 사용하였으며, 나머지는 신경망의 성취도를 평가하는데에 사용하였다. 이중 교육시킨 단어에 대해서는 95%, 교육 받지 않은 단어에 대해서는 85%의 성취도를 나타내고 있다. 음절단위로 변동 규칙의 성취도는 89%를 나타내고 있으며, 또한 종성처리 net에서는 95%의 성취도를 나타낸다. 따라서 종성 처리가 잘 되지 않으면 상위 net의 성취도가 떨어지므로 하위 net의 성취도를 증갈시킬 수 있는 방향으로 해야 한다. 즉 잘못된 데이터가 상위 net로 입력되므로 교육받은 규칙과는 동 떨어질 결과를 나게된다. 음운 규칙중 가장 결과가 잘 나온것은 연음법칙으로 100%의 성취도를 나타냈으며 가장 못배운 규칙은 상호동화로 75%의 성취도를 나타내었다.

한국어는 음절의 음량, 즉 상대적 길이를 잘 나타내야 발화의 흐름이 잘 되어지는 음절-시간 리듬(syllable-timed-rhythm)언어이므로, 연속 음성에서 음절의 지속시간은 매우 중요하다. 따라서 본 연구에서는 문장에서 음절수가 많은 단어들은 상대적으로 음절의 지속 시간을 줄여서 합성한 결과보다 더 자연스러운 합성음을 얻을 수 있었다.

합비음의 청취는 과학적인 인식율 test를 사용하여 음성 합성 시스템의 성능을 평가하여야 하나 본 연구에서는 아직 과학적인 인식율 test를 하지 못하였다. 20대의 7-8명의 성인을 대상으로 하여 듣기 평가를 수행한 결과를 종합해 보면, 'ㅅ', 'ㅈ' 같은 몇몇 변이음에 대해서는 예외가 있었으나 원음성과 비교시 거의 인식할 수 있었다. 또한 미리 합성음에 대해 알려준 경우는 완벽하게 인식하였으나, 그렇지 않은 경우는 합성음에 대한 기대가 커서 완벽하게는 인식할 수 없었다.

LPC에서 음성 시료를 all-pole model로서 vocal track을 modeling 함으로 무성음과 비음이 정확히 모델링 되지 않아 합성음의 음질이 떨어진다. 따라서 여러 응용 분야에서 음성 합성 시스템을 이용하기 위해서는, 그리고 음질의 개선을 위해서는 이들 음소들에 대한 과학적인 연구가 요구된다.

Ⅷ. 결 론

본 논문에서는 합성단위를 반음절로 하여 음성 합성 시스템을 구현하고, 임의의 문장을 한국어 발음기호로 바꾸어 주는 음운 변동 시스템을 신경망 모델을 이용하여 구현하며, 명료도의 향상을 위해 음절 연결규측및 모음간접속 규칙을 제안하며, 음율 조절규칙을 적용한다. 그러나 우수한 음운 변동 시스템및 합성시스템을 구현하기 위해서는 국어에 대한, 특히 무성음이나 비음에 대한 연구및 음소들에 대한 과학적인 연구가 필요하며, 많은 음성 데이터의 분석에 의한 표준 피치 패턴과 지속시간에 대한 연구가 필요하며, 상업적인 분야에 적용하기 위해 H/W에 의한 실시간 음성 합성기의 개발이 요구된다.

參 考 文 獻

[1] 김병수, 윤기선, 박성한 "음절단위를 이용한 한국어 음성합성에 관한 연구" 전자공학회 추계학술대회, 1988.

[2] Fallside, F. and Woods, W.A. "Computer Speech Processing," *Prentice Hall*, 1985.

[3] Bristow, G. "Electronic Speech Synthesis," McGraw Hill, 1984.

[4] Witten, I.H. "Principles of computer Speech," *Academic Press*, 1985.

[5] Rabiner, L.R. and Schater, R.W. "Digital Processing of speech signal," *Prentice Hall*, 1978.

[6] Flanagan, J.L. "Voice of Man and Machine," *JASA*, May 1972.

[7] 허 웅 "국어 음운학," 정음사, 1982.

[8] 이철수 "한국어 음운학," 인하대학교 출판부, 1985.

[9] Akers, G. and Lenning, M. "Intonation in Text-to-Speech Synthesis : evalution of algorithm," *JASA*, vol. 77, Jun 1985.

[10] Sejnowski, T.J and Rosenberg, C.R. "Parallel Networks that Learn to Pronounce English Text," *Complex system 1*, pp. 145-16, 1987

[11] Rumelhart, D. and Hinton, G.E. and R.J. Williams "Learning internal representations by error propagation,": *PDP* vol. 11. MIT. 1986.

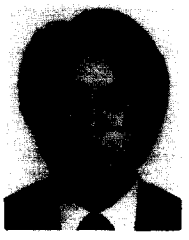
[12] Lippman, R.P "An introduction to computing with Neural Nets," *IEEE ASSP magazine*, 1987, Jan. pp. 4-22.

[13] Markel, J.D. and Gray, A.H. "Linear Prediction of Speech," *Springer-Verlag* 1976.

[14] Pierrehumbert, J. "Synthesizing Intonation,' JASA, vol. 70, Oct. 1981.

[15] Hyun Bok Lee, "Korean Prosdoy: speech rhythm and intonation," *Korea Journal*, vol. 27, no. 2, Feb. 1987.

著 者 紹 介



尹 起 善(正會員)

1966年 3月 25日生. 1988年 2月 한양대학교 전자공학과 공학사 학위취득. 1990年 2月 한양대학교 전자공학과 공학석사 학위취득. 1990年 1月～현재 현대전자 통신개발 1부 근무. 주관심분야는 DSP, 음성합성, 이동통신 등임.

朴 成 漢 (正會員) 第25卷 第12號 參照

현재 한양대학교 전자계산학과 부교수