論文94-31B-1-2

음절을 기반으로한 한국어 음성인식

(Korean Speech Recognition Based on Syllable)

李 漢 鎬*, 丁 弘**

(Young Ho Lee and Hong Jeong)

要 約

인식어휘수를 늘이기 위해서는 기존의 단어단위의 방법으로는 어려움이 있다. 이를 해결하기 위해서는 인식단위를 보다 근본적인 요소인 음절이나 음소로 정해야 한다. 한국어의 경우 음절은 초성,중성,종성으로 구성되어있고 음성으로부터 음절을 구분하기가 쉽다는 특징이 있다. 본 논문에서는 이러한 한국어의 특징을 이용한 음성인식 시스템을 소개하고 있다.

인식알고리즘으로는 시간지연 신경망을 사용하였다. 많은 수의 인식요소를 인식하기 위하여 시스템은 인식대상에 따라 각각 초성,중성,종성인식신경망으로 모듈화하여 구성되었다. 시스템은 먼저 초성,중성,종성을 인식하고 이를 이용하여 고립단어를 인식하게 된다. 실험을 통하여 초성은 2735개의 패턴에 대하여 85.12%, 중성은 3110개의 패턴에 대하여 86.95% 그리고 종성은 1615개의 패턴에 대하여 90.58%의 인식률을 얻었다. 그리고 250개의 고립단어에 대해서는 71.2%의 인식률을 얻었다.

Abstract

For the conventional system based on word, it is very difficult to enlarge the number of vocabulary. To cope with this problem, we must use more fundamental units of speech. For example, syllables and phonemes are such units. Korean speech consists of initial consonants, middle vowels and final consonants and has characteristic that we can obtain syllables from speech easily. In this paper, we show a speech recognition system with the advantage of the syllable characteristics peculiar to the Korean speech.

The algorithm of recognition system is the Time Delay Neural Network. To recognize many recognition units, system consists of initial consonants, middle vowels, and final consonants recognition neural network. At first, our system recognizes initial consonants, middle vowels and final consonants. Then using this results, system recognizes isolated words. Through experiments, we got 85.12% recognition rate for 2735 data of initial consonants, 86.95% recognition rate for 3110 data of middle vowels, and 90.58% recognition rate for 1615 data of final consonants. And we got 71.2% recognition rate for 250 data of isolated words.

*正會員, 三星電子 通信開發室
(Communication R&D Center, Samsung Electronics co. LTD)
**正會員, 浦項工科大學 電子電氣工學科
(DEpt. of Elec. & Electrical Eng., POSTECH)
※이 논문은 1992-1993년도 인공지능연구소 연구비 지원에 의한 것임.
接受日字 : 1993年 1月 21日

Ⅰ. 서론

고도 정보화 시대로 변해감에 따라 각종 정보기기들과의 접촉이 빈번해지고 있다. 컴퓨터가 일상생활에 파고드는 것이 좋은 예라 할 수 있을 것이다. 이에 따라 인간과 기계의 의사소통을 좀 더 자연스러운 방법으로 하고자 하는 욕구는 당연하다 할 수 있다. 즉 인간과 기계와의 대화를 인간과 인간의 대화에 가깝게 하고자 하는 것이다. 이러한 것을 달성하기 위해서는 여러가지 방법이 있겠지만 그중의 하나가 인간의 가장 기본적인 통신수단인 음성을 이용하는 것이다. 음성은 특별한 배움이 없이도 모든 사람들이 사용할 수 있는 통신수단이다. 이러한 음성을 이용한 맨-머신 인터페이스가 이루어지기 위한 핵심적인 기술중의 하나가 음성인식이다. 음성인식이란 인간의 음성을 인식 알고리즘을 통해 단어나 문장으로 전환하는 것이다. 선진국에서는 1950년대부터 음성인식에 관한 연구를 계속해 오고 있다. 국내에서는 아직 연구기간이 짧아서 기초적인 연구를 하고 있다.[15,19,20] 그러나 소용량 인식어휘의 단어인식에 대해서는 이미 실용화 단계에 이르고 있다.[16,17,18]

음성인식에 대한 연구가 진행됨에 따라 음성인식시스템이 해결해야할 문제점중의 하나인 인식어휘수를 늘이는 문제가 제기되고 있다. 인식어휘수를 늘이는데 있어서 기존의 단어 단위의 인식시스템으로는 인식어휘수의 증가에 한계가 있게 된다. 먼저 인식어휘수가 늘어남에 따라 이들을 학습시키기 위한 시간과 이에 필요한 학습패턴이 기하급수적으로 늘어나게 된다. 신경망의 학습시간을 줄이기 위한 연구도 꾸준히 계속되고 있으나[2,6] 인식어휘의 증가에 따른 학습시간의 증가를 감당하기에는 근본적으로 부족하다. 또한 신경망의 출력이 인식하고자 하는 어휘수와 같은 숫자이어야 하므로 신경망의 수용능력에도 한계가 있게 된다.

이러한 문제점의 해결방안으로 인식대상을 단어를 구성하고 있는 보다 근본적인 요소로 정하여 이들을 인식하고 그 결과를 이용하여 최종적으로 단어를 인식하는 방법을 생각할 수 있다. 실제로 단어는 음절과 더 나가서는 음소로 구성되어 있다. 한국어의 경우 2000여개의 음절과 40여개의 음소들로 모든 단어들이 구성되어 있으므로 한정된 수의 이들 요소들을 인식할 수 있다면 무한단어 인식이 가능할 수 있는 것이다. 한국어는 특히 음절이 발달하여 각 음절사이에 비교적 명확한 구분을 지을 수가 있는 특징을 지니고 있다. 본 연구에서는 한국어의 이러한 특징들을 이용하여 음절을 인식단위로 하는 인식시스템을 제안하고 있다. 음절단위의 인식은 음소단위의 인식에 비하여 계산량이 적기 때문에 범용 컴퓨터에서 비교적 빠른 시간안에 단어를 인식할 수 있다.

음절을 인식단위로 하는데에 있어 음절 전체를 하나의 단위로 인식을 하는 것이 좋겠지만 한국어의 경우 2000여개의 음절이 있으므로 이들 모두를 인식하는데는 단어단위의 인식시스템의 경우와 같은 어려움이 생기게 된다. 그러나 한국어의 경우에는 음절구분이 명확할 뿐만아니라 또한 각 음절들은 다시 초성, 중성, 종성의 요소들로 구성되어 있다. 따라서 이들을 이용하고 음절을 인식하고 단어를 인식한다면 대어휘 인식이 가능하게 될 것이다. 본 연구에서는 한국어의 이러한 특징을 이용하여 단어를 인식하는 시스템을 구현하였다.

본 논문에서는 시간지연 신경망을 이용하여 한국어 음성을 초성,중성,종성의 단위로 인식하는 시스템을 설명하고 있다. Ⅱ장에서는 시스템을 구성하는데 사용된 신경망인 시간지연신경망에 대해서 설명하고 Ⅲ장에서는 시스템의 전체 구성을 설명한다. Ⅳ장에서는 음성인식 시스템의 실험 결과를 보여주고 검토를 한다. 마지막으로 Ⅴ장에서 결론을 제시하고 끝을 맺는다.

Ⅱ. 시간지연신경망

신경회로망은 인간의 신경조직을 모방하여 신경조직의 뉴런(neuron)이 하는 기본적인 기능을 수행하는 요소들이 병렬로 상호 연결되어 구성된다. 이에 따라 신경회로망은 병렬처리에 의한 높은 계산속도, 학습능력등을 가지게 되어 음성인식에 있어서 기존의 알고리즘이 갖지 못하는 새로운 가능성을 제시하고 있다. 신경회로망은 인간의 두뇌에 대한 연구가 진척이 됨에 따라 이를 모델링하기 위한 시도로부터 생겨났다. 1949년에 Hebb가 신경회로망 연구의 출발점이 될 학습법칙을 발표했고 50,60년대에는 Minsky, Rosenblatt, Widrow등이 이를 이용하여 퍼셉트론이란 신경회로망을 발표했다. 그러나 1969년에 Minsky와 Papert가 단층 퍼셉트론이 여러가지 단순한 문제들을 해결할 수 없다는 것을 이론적으로 증명하자 연구가 거의 중단되게 되었다. 신경회로망이 음성인식에 사용되게 된 것은 80년대에 신경망이 가지는 단점을 극복할 수 있는 방법이 알려지자 신경망에 대한 관심이 다시 고조되기 시작하면서부터이다. 즉 1969년 이후 많은 기간동안 다층신경망(multi-layer neural network)을 학습시킬 알고리즘이 개발되지 않았었다. 그러다가 Werbos(1974)와

Parker(1982) 그리고 Rumelhart, Hinton, Williams(1986) [10] 등이 각기 다층신경망 학습알고리즘인 역전파알고리즘(Back Propagation)을 발표하였다. 이는 신경회로망에 대한 관심이 다시 고조되게 하는데 결정적인 역할을 하였다.

음성에는 모음의 포만트와 같은 정적인 특성과 음소및 단어의 결합과정에서 나타나는 동적인 특성이 있다. 따라서 음성을 인식하는데는 일반적인 정적구조 신경망에 동적요소(delay,integration)를 첨가한 동적구조 신경망이 좋은 결과를 보이고 있다. 동적구조 신경망에는 다층신경망에 시간지연요소(delay)를 첨가한 시간지연신경망 [4,5,8,9,11,13] 이나 회귀구조를 첨가한 회귀신경망(Recurrent Neural Network) [3] 등이 있다. 여기서는 실험에 사용된 시간지연 신경망의 구조와 학습방법에 대해서 알아본다. 시간지연 신경망은 음성의 동적인 특성을 찾아내기 위해 역전파 신경망에 시간지연요소를 첨가한 것으로 Waibel에 의해 제안되었다. 이는 영어와 일어의 고립단어 인식에 사용되어 높은 인식률을 얻었고, Waibel,Lang 그리고 Hinton [8] 에 의해 알파벳중 가장 혼동하기 쉬운 "B","D","E","V",를 인식하는 실험에 사용되어 좋은 결과를 얻었다.

1. 시간지연신경망의 구조

신경망의 기본단위들은 입력에 연결강도(weight)가 곱해진 합을 구하여 이를 역치(threshold)함수나 시그모이드(sigmoid)함수를 통과시킨후 그 출력을 다음단에 전달하게 된다. 시간지연 신경망에서는 이러한 기본단위들에 시간지연요소를 도입하여 그림 1과 같이 변형한다.

이렇게 시간지연요소를 도입하여 시간지연신경망은 현재의 입력과 과거의 입력을 연관시켜 비교할 수가 있게 된다. 이때 사용하는 함수는 수학적인 편리성때문에 시그모이드 함수를 사용한다.

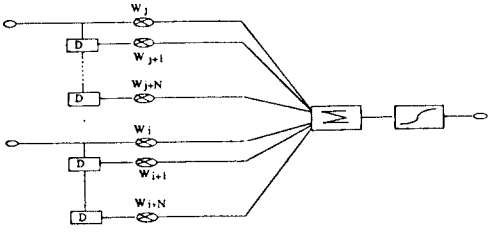

그림 1. TDNN의 구성단위

FIG. 1. Structure unit of TDNN.

전체적으로는 다층인식자(multilayer perceptron)와 같은 세개의 층, 두개의 은닉층(hidden layer)과 출력층으로 구성된다.은닉층에서는 시간지연요소를 통한 입력으로 음성의 국부적인 특징들을 감지해내고 출력층에서는 전단의 은닉층의 시간적인 지연을 갖는 출력의 제곱을 더하여 출력을 하게된다. 이렇게 해서 최종적으로 은닉층에서는 음성의 국부적 특성을 감지함으로써 패턴을 시간적으로 굴곡(time-warping)하게 되고 출력층에서는 전단의 시간적으로 지연된 출력들을 합함으로써 입력패턴에 지연현상이 발생하여도 같은 출력을 낼 수 있는 특징이 생기게 된다. 그림 2에 시간지연신경망의 전체 구조를 나타내었다.

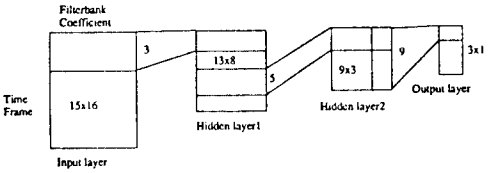

그림 2. 출력이 3개인 TDNN 구조

FIG. 2. Structure of TDNN which has three output nodes.

2. 시간지연신경망의 학습

일반적인 역전파신경망에서는 목적함수가 식(1)과 같이 정의 된다.

$$E=\frac{1}{2}\sum_{j}(O_j-T_j)^2 \tag{1}$$

여기서 출력값 O_j는 j번째 출력노드 y_j의 활성화값(activation value)이다. 이것을 감안하여 식E를 y_j에 대해 편미분하면 출력노드의 활성화값에 대한 오차의 편미분값을 얻게 된다.

$$\frac{\partial E}{\partial y_j}=y_j-d_j \tag{2}$$

이 값이 역전파 학습알고리즘의 역전파 출발점이 된다. 그러나 시간지연신경망에서 출력값은 전단의 은닉층의 지연된 출력의 제곱을 더한 것(3)이므로 이를 E의 정의에 대입하여 y_j에 대해 편미분하면 일반적인 역전파신경망에서와 같이 오차의 편미분값을 얻게 되고 이 값이 시간지연신경망에서의 역전파학습의 출발점이 된다. 은닉층에서는 일반적인 역전파 알고리즘을 사용하여 학습을 하게 된다.에러 역전파시에

연결강도(weight)를 변화시킬때 음성 특성의 시간적 위치에 무관하게 이를 찾아낼수 있도록 시간지연된 연결강도의 평균을 구해 그 값을 각 연결강도에 다시 복사한 다음 에러 역전파를 계속한다.

$$O_i = \sum_t y_{jt}^2 \tag{3}$$

$$\frac{\partial E}{\partial y_{jt}} = 2y_{jt}\left(\left(\sum_t y_{jt}^2\right) - d_j\right) \tag{4}$$

시간지연신경망은 많은 연결강도를 가지고 있으므로 학습에 많은 시간이 걸리게 된다. 따라서 학습을 빨리하기 위해서 고속화 알고리즘을 사용한다.[2,6] 실험에 사용된 방법은 다음과 같다.

· 시그모이드함수를 변형하여 출력을 -1과 1사이로 한다.

$$F(X) = \frac{2}{1+\exp^{-X}} - 1$$

· 출력에러가 기준치 이하의 입력데이타에 대해서는 역전파 학습과정을 생략한다.

· 학습초기에는 입력데이타를 적은 수로 제한하고 시간이 지남에 따라 점차 많은 데이타를 사용한다.

· 에러를 급격히 줄이고 진동을 줄이기 위해 모멘텀(momentum)을 시간에 따라 변화시킨다.

이와 같이 구성된 시간지연 신경망은 많은 응용이 되어 좋은 결과를 얻었다. 그러나 최근에는 단순한 응용보다는 시간지연신경망 자체의 성능을 향상시키려는 연구도 병행되고 있다. 예를 들어 시간축뿐만아니라 주파수축에서도 지연현상을 극복할수 있는 특성(shift invariant)을 갖게하거나[11] 회귀구조를 가지게 하거나[13] 입력패턴의 경계점에서도 정보를 처리할 수 있도록 하는 연구가[12] 진행되고 있다.

Ⅲ. 전체 시스템의 구성

여기서는 한국어 음소의 분류와 신경망의 입력패턴과 학습에 대해서 설명한다. 그리고 인식 알고리즘과 언어처리 알고리즘에 대해서 알아본다. 그림 3에 전체 시스템을 간략하게 나타내었다.

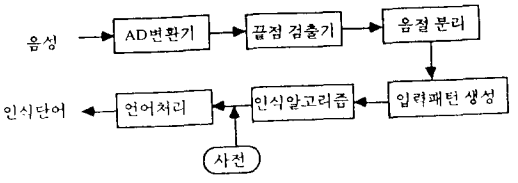

그림 3. 전체 시스템의 구성
Fig. 3. Structure of total system.

1. 한국어 음소의 체계적인 분류

신경망의 구조가 커질수록 이를 학습시키기 위한 시간이 매우 늘어나게 되고 학습시키기 위한 학습데이타의 수도 더욱 많이 필요하게 된다. 더욱 빠른 알고리즘과 하드웨어가 개발되었다고 해도 새로운 종류의 데이타가 생기거나 일부의 데이타가 바뀌었을때 전체 신경망을 다시 학습시키는 것은 비효율적인 일이다. 따라서 신경망을 많은 종류의 데이타에 대해서 학습시키기 위해서는 전체를 작은 모듈들로 나누어 각각의 모듈들을 학습시키는 방법을 사용하여야 할 것이다.[9,11,14] 음절은 초성,중성,종성의 요소들로 구성되어 있으므로 이들의 음성학적 특징을 고려하여 혼동하기 쉬운 몇개의 음소들로 구분한다면 신경망을 효율적으로 이용할 수 있을 것이다. 본 연구에서는 한국어 음소인식에 관한 연구[15] 에서 사용된 음소 구분방식을 변형하여 사용하였다.

표 1. 한국어 초성,종성의 체계적인 분류
Table 1. Classification of Korean initial consonants and final consonants.

	그룹	조음방법	음소
초성	I1	폐쇄음	ㄱ ㄷ ㅂ
	I2	경음	ㄲ ㅃ ㅉ
	I3	격음	ㅋ ㅌ ㅍ
	I4	마찰음	ㅅ ㅆ ㅎ
	I5	파찰음	ㅈ ㅉ ㅊ
	I6	유음(비음)	ㄴ ㄹ ㅁ
종성	F1	폐쇄음	ㄱ ㄷ ㅂ
	F2	유음(비음)	ㄴ ㄹ ㅁ ㅇ

한국어를 이루고 있는 음소는 우선 자음과 모음으로 구분할 수 있으며, 자음은 음절내의 위치에 따라 초성과 종성으로 나뉘어진다. 초성은 조음방법에 따라 폐쇄음,격음,경음,마찰음,파찰음, 그리고 유음으로 구분할수 있다. 종성의 경우는 한국어의 7종성법칙을 이용하여 폐쇄음과 유음으로 구분하였다. 모음의 경우 음을 단독으로 발음할 경우 혀끝의 위치에 따라 FH(Front-High),FL(Front-Low),BH(Back-High),BL(Back-Low)의 4개 그룹으로 구분하였다. 이는 혀끝의 위치에 따라 모음의 중요한 요소인 제1포만트와 제2포만트가 결정되기 때문이다. 이들 포만트는 음소의 파형을 주파수 도면으로 변환하였을때 관측할 수 있는 에너지의 정점에 해당하는 특정 주파수를 말하며, 모음의 경우 이들 포만트의 변화가 적고 중복되지 않기 때문에 일찍부터 음성인식에 도입되어 사용되었다. 표 1과 표 2는 한국어의 음소를 체계적으로 구분한 것이다.

표 2. 한국어 중성의 체계적인 분류
Table 2. Classification of Korean middle vowels.

	그룹	혀끝의 위치	음소
중성	M1	Front-High	아 야 어 여 와
	M2	Front-Low	이 위 외 으
	M3	Back-High	우 유 오 요 워
	M4	Back-Low	에(애) 예(얘) 웨(외 왜)

2. 입력패턴의 생성

마이크로 입력된 음성은 먼저 12kHz의 sampling frequency와 12bit 의 품질을 가지는 디지탈데이타로 변환된뒤 pre-emphasis과정을 거치게 된다. 이는 다음 식(5)를 이용하여 수행되었다.

$$S^*(n) = S(n) - 0.95*S(n-1) \tag{5}$$

표 3. Mel-scale된 주파수 밴드
Table 3. Mel-scaled frequency bands.

	FFT points	Frequency (Hz)		FFT points	Frequency (Hz)
1	0 - 2	0 - 130	9	35 - 41	1673 - 2021
2	3 - 5	131 - 278	10	42 - 49	2022 - 2417
3	6 - 8	279 - 445	11	50 - 58	2418 - 2863
4	9 - 12	446 - 634	12	59 - 69	2864 - 3369
5	13 - 17	635 - 848	13	70 - 81	3370 - 3940
6	18 - 22	849 - 1089	14	82 - 94	3941 - 4586
7	23 - 27	1090 - 1363	15	95 - 109	4587 - 5317
8	28 - 34	1363 - 1672	16	110 - 128	5318 - 6144

이 과정을 거친 음성신호는 매 5msec마다 해밍 윈도우(Hamming window)를 취한 256개의 음성 샘플을 FFT하여 128개의 주파수 성분을 얻는다. 이들 128개의 주파수 성분을 신경망의 입력으로 하기 위해서 멜-스케일된 16개의 주파수대에 존재하는 에너지의 양을 구하게 된다. 멜-스케일된 주파수대는 표 3에 나타내었다. 이때 멜-스케일된 주파수대에 속하는 주파수 성분들은 삼각형 윈도우를 취하여 경계효과를 줄인다. 시간적으로는 매 5msec마다 만들어진 입력 프레임 두개를 평균하여 10msec단위의 프레임을 만들고 음소의 변화과정을 충분히 나타낼 수 있게 하기 위하여 입력패턴의 갯수를 15개로 정하였다.

이렇게 만들어진 입력패턴을 신경망에 사용하기 위하여 신경망의 특성에 맞도록 입력패턴의 각 성분들이 -1과 1사이의 값을 갖도록 정규화 과정을 거치게 된다. 음절단위의 인식시스템에서는 하나의 단어에서 초성,중성,종성들의 입력패턴을 추출하게 되므로 음성에너지의 상대적인 크기도 중요하게 된다. 따라서 하나의 입력패턴에 대해서 정규화하는 것은 의미가 없게 되므로 모든 입력패턴에 대해 다음과 같이 정규화를 시키게 된다.

$$x = +1, \qquad x > 150\text{dB} \tag{6}$$

$$x = -1, \qquad x < 30\text{dB} \tag{7}$$

$$x = \frac{2}{120}(x - 30) - 1, \qquad 30\text{dB} < x < 150\text{dB} \tag{8}$$

그림 4는 이러한 과정을 통해 얻은 입력패턴의 모습을 보이고 있다.

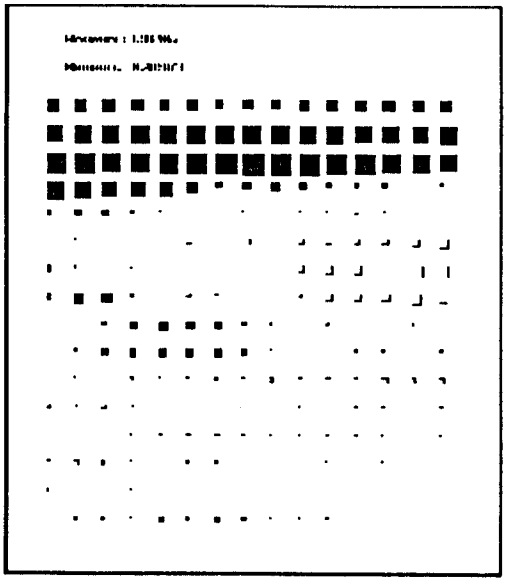

그림 4. 신경망의 입력패턴
Fig. 4. Input pattern of neural network.

3. 신경망의 학습

음성데이타베이스로부터 음성패턴을 추출하여 이를 시간지연신경망 학습 프로그램으로 학습을 하게 된다. 이번 연구에서는 단음절과 단어를 사용하여 음성데이타베이스를 구성하였다. 한 남성화자가 심험실의 환경에서 549개의 단어와 124개의 단음절을 발성하여 12bit 12kHz sampling frequency의 품질을 가지는 디지탈데이타로 컴퓨터에 저장하였다. 이렇게 마련된 음성데이타베이스로부터 표 1과 표 2에서 구분한 요소들에 해당하는 학습패턴들을 만들게 된다. 이때 데이타의 위치에 따른치우침을 없애기 위해 음

소들을 중심에서 뿐만아니라 앞뒤로 10msec씩 이동하면서 학습패턴을 추출하게 된다. 그림 5에서는 음성데이타로부터 학습패턴을 추출하는 예를 보여주고 있다. 본 연구에서는 총 15개의 신경망을 음성데이타베이스로부터 추출한 7555개의 학습데이타로 학습을 시켰다. 15개의 신경망중 3개는 각각 초성,중성,종성에서 그룹을 판단하는 스위치 역할을 하는 신경망이고 나머지 12개의 신경망은 표 1과 표 2에서 구분한 음소들의 그룹들에 속하는 음소를 인식하는 역할을 하게 된다. 먼저 학습데이타들을 각 그룹별로 모아서 각 그룹들의 인식기들을 학습시키게 된다. 각 그룹인식기들의 학습이 끝나면 이들을 다시 모아서 그룹들을 구분하는 스위치 역할의 신경망을 학습시키게 된다.

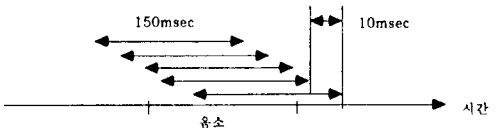

그림 5. 음성데이타로부터 학습패턴 추출의 예
Fig. 5. Extraction of learning pattern from speech data.

학습에 사용한 시간지연신경망은 Waibel이 "B, D, G, V"를 인식하기 위해서 고안한 것을 출력노드의 수를 고려하여 적절하게 변형하여 사용하였다. 표 4에서 출력수에 따른 시간지연신경망의 구조를 보이고 있다. 각각의 시간지연신경망을 학습시키는 과정은 고속화 알고리즘을 사용하지만 시간지연신경망 자체가 방대한 연결강도를 가지고 있으므로 전체 시스템을 학습시키는데 약 20MIPS의 성능의 컴퓨터에서 2주일 가량이 소요되었다.

표 4. 출력수에 따른 TDNN의 구조
Table 4. Structure of TDNN according to the number of ouput nodes.

Output number	Input layer	Hidden layer1	Hidden layer2
2	16x15	6x13	2x9
3	16x15	8x13	3x9
4	16x15	10x13	4x9
5	16x15	12x13	5x9
6	16x15	15x13	6x9

4. 인식시스템의 구성

전체 시스템의 학습이 끝나면 이를 통하여 얻어지는 신경망의 연결강도를 이용하여 주어진 음성을 인식할수 있는 인식시스템을 구성하게 된다.

음성을 마이크로 발음하면 이는 AD변환기를 통해 12bit 12kHz로 디지탈화하여 컴퓨터에 입력이 된다. 이 데이타는 끝점 검출기와 음절분리 알고리즘을 거치게 되어 음절의 갯수와 각 음절의 모음의 위치를 알아내게 된다. 알려진 모음의 위치를 중심으로 앞뒤로 초성과 종성 자음의 위치를 추측하여 각각 초성, 중성, 종성 인식신경망에 입력을 하게 된다.

1) 끝점 검출기

배경잡음으로부터 음성부분을 분리해내는 작업을 끝점검출(Endpoint detection)이라 한다. 고립단어 인식시스템에 있어서는 정확한 끝점 검출은 다음의 두가지 이유에서 중요하다. [1,7]

· 인식시스템의 인식률은 정확한 끝점 검출에 영향을 받는다.

· 인식과정의 계산량은 정확한 끝점 검출을 통해 최소화 될수 있다.

본 연구에서는 음절분리 알고리즘을 적용하기 전에 계산량을 최소화하기 위하여 끝점 검출기를 도입하였다. 본 연구에서 구현된 끝점 검출기는 adaptive level equalizer와 energy pulse detector로 구성되어있고 단시간에너지(short-time energy)와 영교차율(zero crossing rate)을 이용하여 설계하였다.

$$\text{Short time energy } E(n) = \sum_{k=n-N+1}^{n} x(k)^2 -$$

$$\text{Zero crossing rate } Z(n) = \sum_{k=n-N+1}^{n} |x(k) - x(k-1)|$$

여기서 x(n)은 디지탈화된 음성 샘플을 의미한다. N은 단시간에너지(short-time energy)와 영교차율을 구하기 위한 윈도우의 크기로 실험에서는 20msec의 길이로 정하고 이를 10msec단위로 이동하면서 단시간에너지(short-time energy)와 영교차율을 구하였다.

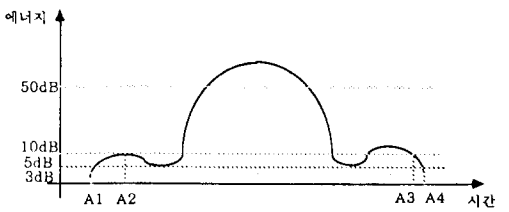

그림 6. 끝점검출기에 사용되는 문턱값
Fig. 6. Thresholds used in endpoint detector.

Adaptive level equalizer

E(n)과 Z(n)은 그 값이 입력 음성마다 달라지게 된다. 따라서 다음단의 energy pulse detector에 사용되는 문턱값(threshold)의 절대값을 구하기 위해서는 이들 단시간에너지(short-time energy)와 영교차율은 배경잡음에 따라 정규화되어야 한다. 이때 사용되는 식은 아래와 같다.

$$E(n)^* = 20*\log_{10}E(n) - Q_e$$
$$Z(n)^* = 20*\log_{10}Z(n) - Q_z$$

여기서 Qe와 Qz는 배경잡음으로부터 얻은 평균값이다. 이렇게 정규화된 단시간에너지(short-time energy)와 영교차율은 음성이 아닌곳에서는 0dB를 중심으로 변하는 값이 되고 음성이 있는 곳에서는 상당히 큰 값을 갖게 된다.

Energy pulse detector

Adaptive level equalizer를 통과하여 나온 E(z)를 이용하여 에너지 펄스를 구하게 된다. 여기서 사용되는 문턱값들을 그림 6에 나타내었다. 음성신호를 왼쪽에서 오른쪽으로 이동하면서 에너지 펄스를 찾게 된다.에너지의 값이 3dB를 넘는 점이 생기면 이를 기록하고 에너지가 다시 10dB를 넘게되면 처음 3dB를 넘었던 점을 에너지 펄스의 시작점으로 한다. 단 10dB를 넘는점과 3dB를 넘는 점의 간격이 길 경우 10 dB를 넘는 점을 시작점으로 한다. 에너지 펄스의 끝점을 찾는 과정도 이와 같다. 단지 문턱값을 3dB 대신 5dB로 한다. 에너지 펄스를 찾게 되면 이들 펄스의 최고값과 지속기간을 검사한다. 음성일 경우 에너지가 50dB를 넘고 지속기간은 80msec를 넘는다고 가정하고 이에 미치지 못하는 에너지 펄스는 잡음으로 간주하여 버리게 된다.

다음으로 마찰음과 같이 에너지는 작고 영교차율이 큰 자음들의 경우를 고려하여 보다 정확하게 구별하기 위하여 에너지 펄스의 시작점에서 영교차율이 3dB를 넘을 경우 시간을 거슬러가면서 영교차율이 다시 3dB보다 작아지는 점을 찾아 그점을 단어의 시작위치로 결정한다.

2) 음절분리 알고리즘

음절에는 하나의 모음이 반드시 들어있다. 이들 모음은 자음에 비해 상대적으로 큰 에너지를 유지한다. 따라서 단어의 에너지 곡선은 음절의 수에 따른 에너지 펄스를 나타내게 된다. 그러므로 에너지를 이용하면 대략적인 모음의 위치와 음절의 갯수를 알 수 있게 된다. 끝점검출기에서와 같이 adaptive level equalizer를 통과하여 나온 E(z)를 이용하여 모음부분을 찾게 된다. 에너지의 값이 40dB를 넘는 점이 생기면 이를 모음부분의 시작으로 한다. 다시 40dB를 내려가는 점이 나오면 하나의 음절의 모음부분을 찾게 되는 것이다. 여기서 구간의 지속기간이 60msec를 넘지 않으면 잡음으로 생각하여 모음구간에서 제외하게 된다. 그림 7에서 에너지를 이용하여 모음구간을 추정하는 것을 보이고 있다.

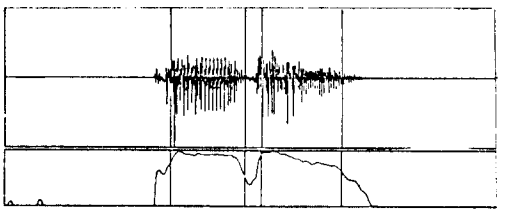

그림 7. 에너지를 이용한 음절의 모음구간 측정
Fig. 7. Vowel interval estimation using short time energy.

에너지를 이용하여 대략적으로 모음의 위치를 파악한 뒤 모음의 위치를 더욱 정확히 측정하기 위하여 모음을 찾기위한 음절구분 시간지연신경망을 설계하고 학습시켰다. 여기서는 음소단위를 구분하는 것이 아니고 그보다는 긴 단위인 자음 모음을 구분하게 되는 것이므로 하나의 시간 프레임을 40msec로 하여 이러한 긴 지속시간을 갖는 데이타를 찾아내도록 설계하였다. 전체 프레임의 갯수는 3개로 하여 모두 120msec의 음성데이타를 입력받도록 하였다. 출력노드의 수는 입력된 음성이 모음인지 아니지를 판별하도록 2개로 설정하였다. 즉 모음이 아닐 경우는 출력(0)가 출력(1)보다 큰 값을 갖게 되고 입력이 모음일 경우는 출력(1)이 출력(0)보다 큰 값을 갖게 된다. 그러나 실제로는 신경망의 출력이 모음이 아닌 부분에서도 출력(1)이 출력(0)보다 큰 값을 갖게 되는 경우가 생기게 된다. 이러한 경우를 보정하기 위하여 모음이 자음에 비해 에너지가 크다는 특징을 이용한다. 즉 시간지연신경망의 출력(1)이 출력(0)보다 커지게 되는 구간에서는 이들 구간의 에너지를 모두 더한후 구간의 길이로 나눈 정규화한 에너지를 비교하여 이값이 문턱값을 넘지 않는 구간은 모음이 아니라고 판정을 하게 된다.

$$\text{Nomalized Energy} = \sum_{i=\text{vowel estimated interval start}}^{\text{vowel estimated interval end}} \frac{\text{Energy}}{\text{Interval length}}$$

Vowel interval if Normalized Energy ≥ 0.4

Not vowel interval, otherwise

이렇게 되면 에너지가 작은 자음 구간에서는 시간 지연신경망이 오동작하여 출력(1)의 값을 크게 하더라도 정규화한 에너지가 작은 값이 되어 모음구간의 추정에서는 제외되게 된다. 그림 8은 알고리즘이 동작하는 것을 보이고 있다.

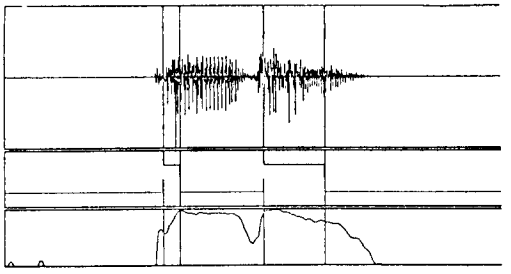

그림 8. 신경망을 이용한 음절의 모음구간 추정
Fig. 8. Vowel interval estimation using neural network.

일단 모음의 위치를 알아내면 모음의 시작과 끝부분에 150msec의 구간을 초성과 종성의 자음위치로 추정하게 된다.

3) 음절의 인식

인식시스템은 먼저 입력된 음성의 그룹을 판별하는 신경망이 동작을 한다. 초성의 경우 표 1에서 구분한 6개의 그룹을 판별하기 위하여 출력노드의 갯수가 6개인 신경망이 동작을 하게 되고 중성의 경우 출력노드가 4개인 신경망이 그리고 종성의 경우는 출력노드가 2개인 신경망이 동작을 하게 된다. 이들 신경망이 동작을 하게 되면 입력패턴이 속하는 그룹을 알수 있게 되고 이로부터 다시 각각의 그룹에 해당하는 신경망이 동작을 하게 된다. 그림 9에서 음절이 인식되는 예를 보이고 있다. 그림 10에서는 음절인식 알고리즘의 구조를 자세히 보이고 있다.

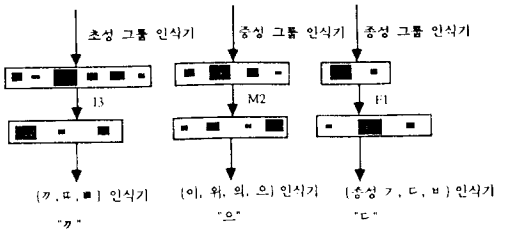

그림 9. 음절을 인식하는 과정
Fig. 9. Syllable recognition.

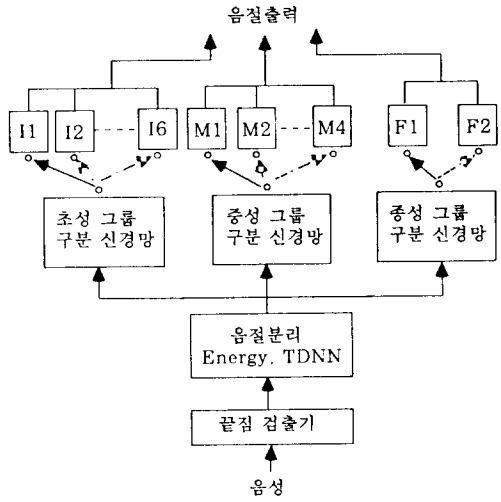

그림 10. 음절인식 알고리즘
Fig. 10. Syllable recognition algorithm.

5. 언어처리 알고리즘

일반적인 음성인식 시스템에서 언어에 따른 제약조건이나 지식등을 이용하여 인식어휘를 정하는 단계가 언어처리 단계이다. 인식어휘수가 적은 음성인식 시스템에서는 언어처리 알고리즘이 그리 중요하지 않다. 그러나 인식어휘수가 늘어남에 따라 많은 단어들이 비슷한 음소들을 가지게 된다. 따라서 이들을 인식알고리즘에서 모두 구별하기는 힘들게 된다. 언어처리 알고리즘에서는 각종 문법지식등을 이용하여 어들이 수록되어 있는 사전을 통해 최종 인식어휘를 결정한다.

본 연구에서는 간단하게 몇가지 규칙을 정하여 이를 가장 만족하는 단어를 인식어휘로 결정하도록 하였다. 앞으로 각종 지식을 이용하는 언어처리기를 연구한다면 시스템의 성능을 크게 높일수 있을 것이다.

음성인식 시스템으로부터는 단어의 초성,중성,종성의 후보들이 출력이 된다. 이들 후보들을 조합하여 가장 적합한 단어를 최종적으로 선택하는 것이 언어처리 알고리즘의 목적이다. 단어사전에는 인식어휘의 음절수와 발음표기에 따른 각 음절의 초성,중성,종성을 기록하고 있다. 언어처리부에서는 이 단어사전을 이용하여 최종 인식어휘를 찾아내게 된다.

먼저 출력된 음절수와 단어사전의 음절수를 비교하여 먼저 음절수가 다른 후보들을 배제하게 된다. 다음에는 인식된 초성,중성,종성의 데이타와 단어사전의 초성,중성,종성 데이타를 비교하여 가장 에러가 적은 것을 인식단어의 후보로 선택하게 된다.

$$X = \min_{i} |D_k(i) - P_k| \quad i = 1,2,\ldots,\text{Maxword}$$

$$k = 1,2,\ldots,n(\text{syllablenumber})$$

D : 단어사전의 수록 어휘

P : 인식된 초성,중성,종성 열

다음에 신경망의 가장 큰 값을 가지는 출력이 오류를 범하였을때 그 다음으로 큰 값을 갖는 출력이 옳게 선택하는 경우가 많으므로 이를 고려하여 최대치 뿐만 아니라 그 다음 값도 후보 선택에 고려되도록 하였다. 즉 초성,중성,종성마다 두개씩의 결과를 내어 이들의 조합으로 단어를 만들어 윗단계에서 찾아진 후보들과 비교한다. 여기서 가장 적은 에러를 갖는 단어가 여럿일 경우 표 5, 7, 9의 결과를 참고하여 쉽게 혼란이 되는 결과들을 보정해 주기 위한 규칙들을 적용하여 다시 후보를 선별하게 된다. 이러한 규칙들을 차례대로 적용하여 가장 적은 에러를 가지는 단어를 인식어휘로 선정하게 된다.

본 논문에서 제안하는 음성 인식시스템은 음절단위로 단어를 인식하게 되므로 한번 학습을 마치게 되면 단어서전을 바꿈에 따라 인식어휘를 교체할 수 있게 된다. 기존의 단어단위의 인식방법에 있어서는 단어단위로 학습을 시키므로 단어가 달라지면 새로 음성을 녹음하여 다시 신경망을 학습시켜야 했다. 따라서 실제 인식 시스템으로 활용하는데 훨씬 유리하다 할 수 있겠다.

Ⅳ. 실험 및 결과

먼저 인식시스템의 성능에 가장 영향을 미치는 초성,중성,종성의 인식률을 알아보고 다음으로 단어의 인식 결과를 알아본다.

1. 음절 구성요소의 인식결과

음절 단위의 인식시스템이 동작을 하기 위해서는 초성, 중성, 종성의 인식신경망을 학습시켜야 하고 이것의 결과가 최종적으로 단어를 인식하는데 있어서 결정적인 역할을 하게 된다. 이들의 인식결과를 표 6, 표 8, 표 10에 나타내었다.

초성의 경우 총 2735개의 패턴에 대해서 85.12%의 인식률을 나타내었고 그룹들을 인식하는데는 88.08%의 인식률을 나타내었다. 중성의 경우 3110개의 패턴에 대해서 86.95%의 인식률을 나타내었고 93.60%의 그룹인식률을 나타내었다. 종성은 1615개의 패턴에 대해 90.58%의 인식률과 98.57%의 그룹인식률을 보였다. 표 5, 표 7, 표9에서는 음들이 오인식되는 경향을 알아볼수 있도록 confusion matrix를 나타내었다.

표 5. 초성의 confusion matrix

Table 5. Confusion matrix of initial consonants.

	ㄱ	ㄷ	ㅂ	ㅋ	ㅌ	ㅍ	ㄲ	ㄸ	ㅃ	ㅅ	ㅆ	ㅎ	ㅈ	ㅊ	ㅉ	ㄴ	ㅁ	ㅇ
ㄱ	339	6	35	2	0	0	19	0	0	9	0	3	13	0	1	8	11	19
ㄷ	2	184	8	0	0	0	2	17	2	0	0	0	0	0	0	2	1	2
ㅂ	0	0	148	0	1	10	3	0	2	0	0	0	0	0	0	3	5	13
ㅋ	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ㅌ	0	0	0	0	85	0	0	0	0	0	0	0	0	0	0	0	0	0
ㅍ	0	0	0	0	0	110	0	0	0	0	0	0	0	0	0	0	0	0
ㄲ	0	0	0	0	0	0	55	0	0	0	0	0	0	0	0	0	0	0
ㄸ	0	0	0	0	0	0	0	75	0	0	0	0	0	0	0	0	0	0
ㅃ	0	0	6	0	0	0	4	4	47	0	0	0	0	0	0	0	0	0
ㅅ	4	0	5	0	2	0	0	0	0	301	0	6	3	0	12	0	0	2
ㅆ	0	0	0	0	0	0	0	0	0	0	55	0	0	0	0	0	0	0
ㅎ	5	5	3	7	27	42	3	0	4	5	0	105	0	0	0	2	2	4
ㅈ	3	5	3	0	0	0	0	0	0	16	0	0	210	1	0	0	1	1
ㅊ	0	0	0	0	0	0	0	0	0	0	0	0	0	55	0	0	0	0
ㅉ	0	0	0	0	2	0	0	0	0	0	0	2	0	0	157	0	1	0
ㄴ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	110	0	0
ㄹ	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	108	3
ㅁ	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	154

표 6. 초성의 인식결과

Table 6. Recognition results of initial consonants.

음소	그룹인식에러수	그룹인식률(%)	초성인식에러수	초성인식률(%)	패턴수
ㄱ	85	81.73	126	72.91	465
ㄷ	26	88.19	36	83.64	220
ㅂ	37	80	37	80	185
ㅋ	0	100	0	100	30
ㅌ	0	100	0	100	85
ㅍ	0	100	0	100	110
ㄲ	0	100	0	100	55
ㄸ	0	100	0	100	75
ㅃ	0	100	8	85.46	55
ㅅ	27	91.65	34	89.86	335
ㅆ	0	100	0	100	55
ㅎ	105	51.17	110	48.84	215
ㅈ	29	87.92	30	87.5	240
ㅊ	0	100	0	100	55
ㅉ	13	92.36	13	92.36	170
ㄴ	0	100	0	100	110
ㄹ	2	98.27	7	93.92	115
ㅁ	1	99.38	6	96.25	160
합계	326	88.08	407	85.12	2735

표 7. 중성의 confusion matrix

Table 7. Confusion matrix of middle vowels.

	아	야	어	여	와	이	위	외	우	유	오	요	워	에	예	네
아	639	6	5	6	19	0	0	0	1	0	1	2	0	6	13	2	0
야	1	49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
어	6	0	214	33	0	0	0	0	11	0	0	16	0	43	2	0	0
여	0	0	0	146	0	0	0	0	0	0	0	1	0	8	0	0	0
와	0	0	0	0	57	0	0	0	0	0	0	0	0	3	0	0	0
이	0	0	0	0	0	288	48	0	11	0	0	4	6	0	4	2	2
위	0	0	0	0	0	2	58	0	0	0	0	0	0	0	0	0	0
외	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0
우	0	0	0	0	0	0	0	0	131	20	1	6	26	1	0	0	0
유	0	0	0	0	0	0	0	0	0	266	0	10	7	0	0	0	0
오	0	0	0	0	0	8	0	0	0	0	45	0	0	0	0	0	0
요	0	0	0	0	0	0	0	0	0	23	0	309	8	0	0	0	0
워	0	0	0	0	0	0	0	0	0	0	0	0	65	0	0	0	0
에	0	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0
예	0	0	0	0	0	0	0	0	0	0	0	0	0	0	252	21	1
네	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	80	0
															0	0	45

표 8. 중성의 인식결과
Table 8. Recognition results of middle vowels.

음소	그룹인식에러수	그룹인식률(%)	중성인식에러수	중성인식률(%)	패턴수
ㅏ	25	96.43	61	91.29	700
ㅑ	0	100	1	98	50
ㅓ	72	77.85	111	65.85	325
ㅕ	9	94.2	9	94.2	155
ㅘ	3	95	3	95	60
ㅣ	18	95.07	77	78.91	365
ㅞ	0	100	2	96.67	60
ㅚ	0	100	0	100	30
ㅗ	54	71.82	54	71.82	185
ㅜ	22	92.79	39	87.22	305
ㅠ	0	100	0	100	45
ㅡ	0	100	31	90.89	340
ㅢ	0	100	0	100	65
ㅟ	0	100	0	100	30
ㅔ	1	99.64	23	91.64	275
ㅖ	0	100	0	100	80
ㅙ	0	100	0	100	45
합계	199	93.60	406	86.95	3110

표 9. 종성의 confusion matrix
Table 9. Confusion matrix of final consonants.

	ㄱ	ㄷ	ㅂ	ㄴ	ㄹ	ㅁ	ㅇ
ㄱ	304	0	6	0	0	0	0
ㄷ	0	45	0	0	0	0	0
ㅂ	1	0	59	0	0	0	0
ㄴ	3	0	0	278	0	10	14
ㄹ	6	11	0	6	208	9	5
ㅁ	0	0	0	21	0	132	12
ㅇ	0	0	3	37	0	8	437

표 10. 종성의 인식결과
Table 10. Recognition results of final consonants.

음소	그룹인식 에러수	그룹인식률(%)	종성인식에러수	종성인식률(%)	패턴수
ㄱ	0	100	6	98.07	310
ㄷ	0	100	0	100	45
ㅂ	0	100	1	98.34	60
ㄴ	3	99.02	27	91.15	305
ㄹ	17	93.07	37	84.99	245
ㅁ	0	100	33	80	165
ㅇ	3	99.39	48	90.11	485
합계	23	98.57	152	90.58	1615

초성의 경우 "ㅎ"가 인식률이 가장 낮았고 중성의 경우는 "어","으","이" 그리고 종성의 경우는 "ㄹ","ㅁ","ㅇ"이 인식률이 낮은 음으로 나타났다.

초성의 경우 인식률이 낮은 음들은 그룹을 인식하는 과정에서부터 오인식이 많이 생김을 알수 있다. 초성의 "ㅎ"의 경우 마찰음이면서도 음이 파열이 되는 성격을 띠고 있어서 경음 그룹과 혼동이 됨을 표 5에서 알수 있다. 중성의 "어","으"의 경우도 단어내에서 발음될 경우 초성과 종성의 영향을 받아 자기의 음가를 제대로 나타내지 못하기 때문에 이러한 패턴들이 인식률에 영향을 미침을 알수 있다. "이"의 경우 초성의 뒤에서 발음될 경우 조음결합현상에 의해 생기는 포만트의 변화가 "위"의 포만트의 변화와 비슷하여 이와 혼동됨을 표 7에서 알수 있다. 종성의 경우 "ㄹ","ㅁ","ㅇ"의 음들이 그룹내에서 오인식이 됨을 볼수 있다. 그러나 "ㄹ"의 경우 다른 음들과는 달리 그룹인식에서도 오인식이 많이 나타나는데 이는 다른 음들보다는 유음의 성격을 덜 띠고 있고 지속시간도 짧은데서 기인하는 것으로 보인다. 이상의 결과를 보면 단어내에서의 조음 결합 현상때문에 음성의 특성의 변하여 이를 인식하는데 어려움을 주는 것을 알 수 있다.

2. 단어인식 결과

인식시스템의 최종성능을 평가하기 위하여 먼저 250개의 단어를 설정하고 이를 한화자가 4회발음하여 얻은 음성신호로 시험하여 결과를 얻었다. 250개의 단어들은 국어사전에서 임의로 선택한 단어들이다. 전체적인 성능을 표 11에 나타내었다. 음절수가 증가함에 따라 인식어휘를 선택하는데 대한 정보량이 많아지므로 인식률이 증가함을 볼수 있다. 그러나 각 음절마다의 초성, 중성, 종성을 인식하게 되므로 음절이 많아질수록 인식대상이 많아져 인식하는데 걸리는 시간이 늘어나게 된다. 세그멘테이션이 인식률에 미치는 영향을 알아보기 위하여 수작업으로 세그멘테이션을 한 음성과 알고리즘에 의해 세그멘테이션이 된 음성의 인식률을 비교하여 보았다. 인식단어들중 인식률이 나쁜 55개의 단어에 대해서 실험을 행하였다. 표 12에서 보면 수작업으로 한 쪽이 인식률이 높음을 알 수 있다. 이로부터 음절분리 알고리즘에서 세그멘테이션이 부정확하여 인식률을 저하시킴을 알 수 있다. 그림 11에서는 실제로 단어가 인식되어지는 과정을 덤프하여 나타내었다.

표 11. 단어인식 결과
Table 11. Recognition results of words.

	인식률	
1음절단어	66/108	61.1%
2음절단어	466/652	71.5%
3음절단어	180/240	75.0%
합계	712/1000	71.2%

표 12. 세그멘테이션 방법에 따른 인식률의 변화
Table 12. Variation of recognition rateaccording to segmentation methods.

	인식률	
알고리즘	23/55	41.8%
수작업	33/55	66%

```
Group 0 0.996 Group2 1 0.304 Index 0 0.8441 Index2 2 0.4562 ㄱ
Group 0 0.302 Group2 3 0.255 Index 0 0.7445 Index2 1 0.1927 아
Group 1 1.397 Group2 0 0.011 Index 0 0.3755 Index2 1 0.1398 ㄴ2
간
Group 4 1.007 Group2 0 0.474 Index 0 0.9115 Index2 1 0.2648 ㅈ
Group 0 1.321 Group2 1 0.138 Index 0 0.3934 Index2 4 0.3126 아
Group 1 0.442 Group2 0 0.365 Index 3 0.5313 Index2 0 0.4799 ㅇ2
장
Final Output : (2) 간장
True Word is 간장
Recognition is True
```

그림 11. 단어인식과정
Fig. 11. Word recognition.

Ⅴ. 결론

본 연구에서는 대용량 단어인식시스템에 사용할 수 있는 음절을 인식단위로 하는 음성인식 시스템을 구현하여 실험하였다. 먼저 12bit 12kHz의 품질로 변환된 음성으로부터 시간지연신경망과 에너지곡선을 이용하여 음절의 모음부분을 찾아내고 이로부터 한국어의 특징인 초성, 중성, 종성의 음절 결합원리를 이용하여 초성과 종성의 위치를 추측하여 이를 각각의 인식 시간지연신경망을 이용하여 인식한 후 그 결과를 미리 준비된 사전과 비교하여 최종적으로 단어를 인식하도록 시스템을 구성하였다. 음소단위의 인식시스템에서는 음소열이 출력되므로 이를 해석하는데 어려움이 있고, 전체 음성을 인식하는데 많은 시간이 걸리게 된다. 그러나 본 연구에서는 한국어의 특징인 음절의 명확성을 이용하여 계산량을 많이 줄일수가 있고 최종적으로 인식단어를 찾아내기 쉽도록 하였다.

그 결과를 보면 화자종속 고립단어 250개에 대하여 71.2%의 인식률을 나타내었다. 음절의 세그멘테이션을 수작업으로 하여 단어를 인식한 실험으로부터 단어의 인식결과가 음절의 초성, 중성, 종성의 위치를 찾는 세그멘테이션 알고리즘의 부정확함에 영향을 많이 받음을 알 수 있다. 이로부터 신경망이 지연현상을 극복하는 특성을 가지고 있다하더라도 세그멘테이션이 인식률에 많은 영향을 미침을 알 수 있다. 세그멘테이션과 후처리, 언어처리부분을 개선한다면 시스템의 성능을 더욱 높일 수 있을 것이다.

參 考 文 獻

[1] E. S. Dermatas and N. D. Fakotakis and K. Kokkinakis, "Fast Endpoint Detection Algorithm for Isolated Word Recognition in Office Environment," *Proc. of ICASSP*, vol. 1, pp. 733-736, 1991

[2] S. E. Fahlman, "Faster-Learning Variations on Back-Propagation: An Empirical Study," *Proc. of Connectist Models Summer School*, 1988.

[3] Fabio Greco, Andrea Paoloni and Giacomo Ravaioli, "A Recurrent Time-Delay Neural Network for Improved Phoneme Recognition," *Proc. of IEEE ICASSP*, pp. 81-84, S2.11 1911.

[4] N. Hataoka and A. Waibel, "Speaker-Independent Phoneme Recognition on TIMIT Database Using Integrated Time-Delay Neural Networks," *Proceedings of IJCNN*, vol. 1, pp. 57-62, San Diego, California, Jun 1990.

[5] A. Hirai and A. Waibel, "Phoneme-Based Word Recognition by Neural Network - A Step Toward Large Vocabulary Recognition," *Proc. of IJCNN*, vol. 3, pp. 671-676, San Diego, California, Jun 1990.

[6] Robert A. Jacobs, "Increase Rates of Convergence Through Learning Rate Adaptation," *Neural Networks*, vol. 1, pp. 295-307, 1988.

[7] L. Lamel, "An Improved Endpoint Detector for Isolated Word Recognition," *IEEE Trans.* on ASSP, vol. ASSP-29, no. 4, Aug 1981.

[8] K. J Lang and A. Waibel, "A Time-Delay Neural Network Architecture for Isolated Word Recognition," *Neural Networks*, vol. 3, pp. 23-43, 1990.

[9] M. Miyatake, H. Sawai, Y. Minami

and K. Shikano, "Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time-Delay Neural Networks," *Proc. of ICASSP*, vol. 1, pp. 449-452, 1990.

[10] D. E. Rumelhart and J. L. McCelland (Eds), Parallel Distributed Processing : *Exploration in the Micro-Structure of Cognition*, vol. 1, MIT Press, pp. 318-362, 1986.

[11] H. Sawai, A. Waibel, P. Haffner, M. Miyatake and K. Shikano, "Parallelism, Hierarchy and Scaling in Time-Delay Neural Networks for Spotting Japanese Phonemes/CV - Syllables," *Proceedings of IJCNN*, vol. 2, pp. 81-88, Washington D.C. Jun. 1989.

[12] Jun-ichi Takami and Shigeki Sagayama, "A Pairwise Discriminant Approach to Robust Phoneme Recognition by Time-Delay Neural Networks," *Proc. of IEEE ICASSP*, pp. 89-92, S2. 13 1991.

[13] A. Waibel, T. Hanazwa, G. Hinton, K. Shikano and K. Lang, "Phoneme Recognition Using Time-Delay Neural Network," *IEEE Trans. on ASSP*, vol. 1 ASSP-37, Mar. 1989.

[14] Alexander Waibel, Hidefumi Sawai and Kiyohiro Shikano, "Modularity and Scaling in Large Phonemic Neural Networks," *IEEE Trans. on ASSP*, vol. 37, no. 12, pp. 1988-1898, Dec 1989.

[15] 정차균, "TDNN을 이용한 한국어 음성인식에 관한 연구," 석사학위논문, 포항공과대학, 1992.

[16] 정차균, 이영호, 최동준, 정홍, "음성명령을 위한 단어 인식시스템," 신호처리 학술대회 논문집, 1991년 9월, pp. 272-275.

[17] 최동준, "로보트 명령을 위한 음성인식/합성시스템," 석사학위논문, 포항공과대학, 1993.

[18] 최동준, 강성호, 정홍, 홍기상, "음성명령과 비젼시스템을 이용한 지능형 로보트환경 구축," 신호처리 학술대회 논문집, 1992년 9월.

[19] "대어휘 연속음성 인식을 위한 음소 인식 기술 개발," 과학기술처, 1991.

[20] "음성명령환경에 관한 연구," 한국전자통신연구소 연구보고서, 1992.

著者紹介

李 漢 鎬(正會員)

1991年 2月 한국과학기술원 과학기술대학 전기및 전자공학과 졸업(학사). 1993年 2月 포항공과대학 전자전기공학과 졸업(석사). 현재 ~ 삼성전자 재직 주관심 분야는 신경회로망을 이용한 음성인식 등임.

丁 弘(正會員) 第 29卷 B編 第 9號 參照

현재 포항공과대학 전자전기공학과 교수