論 文

역전파 학습 신경망을 이용한 고립 단어 인식시스템에 관한 연구

正會員 金 重 泰*

A Study on the Isolated Word Recognition System Using a Neural Network

Jung Tae KIM* *Regular Member*

要 約 본 논문은 음성신호의 실시간 저장법과 기존 표본 데이터에서 개선된 표본 데이터 방법을 제안하여, 신경회로망의 역전파 학습 알고리즘을 이용한 고립 단어 음성인식 시스템에 대하여 연구하였다. 각 층의 노드 수 변화에 의한 기존 표본 데이터방식과 새로운 표본 데이터 방식에서의 인식률과 에러율 변화를 비교하였다. 본 연구 결과, 인식률은 95.1%를 얻었다.

ABSTRACT This paper proposed a real-time memory storage method and an improved sample data method from given data of the speech signal. so, the isolated word recognition system using a back-propagation learning algorithm of the neural network is studied. The recognition rate and the error rate are compared with the new sample data sets generated from small sets of given sample data by the node nunber variation of each layer. In this result, the recognition rate of 95.1% was achived.

I. 서 론

신경회로망은 인간의 생체 구조와 기능에 대한 연구에서 발견된 인간의 뇌와 유사한 모델 방식으로, 기존의 폰 노이만(von neumann) 방식과는 다른 병렬 처리기법을 이용한 새로운 응용 방식이다.

신경회로망은 1960년대 말 Minsky등에 의한 단층 인식자(single-layer perceptron) 모델의 한계성으로 1980년대 까지는 소수 연구자에 의하여 연구되었다.

1982년 HOPFIELD 교수는 일정한 표준치를 기억하였다가 어떤 표본치가 입력될 때 유사한 표본치를 찾아내는 연산기억 방식의 알고리즘을 발표함으로써 신경회로망 연구의 새로운 계기가

*崇實大學校電子工學科
Dept. of Elec. Eng. Soong Sil Univ.
論文番號：90-75(接受 1990. 7. 25)

되었다.

신경회로망에 관한 연구가 진행됨으로 인하여 음성인식, 영상인식, 반도체, 제어, 광학등의 분야에서 새로운 가능성을 보였다[1 2].

음성인식 분야에서 기존 DTW(dynamic time waiping), VQ(vector quantization), HMM(hidden markov models)등의 음성인식 알고리즘은 특정화자 고립단어 인식에서는 높은 인식률을 얻을 수 있으나, 독립화자 음성 인식에서는 개인차 및 음소, 음절, 단어 사이의 특징 추출이 어려워 높은 인식률을 얻기가 곤란하다.

최근의 신경회로망은 종래의 음성인식 알고리즘 보다는, 학습함으로써 경험적 정보 지식의 축적으로 높은 인식률을 얻을 수 있다.

그러나 신경회로망을 이용한 음성인식 분야는 GOLD, BENGGIO, KOHNON 등에 의하여 많은 연구를 하고 있다[3 4].

본 논문은 학습함으로써, 프로그램이 필요없고 새로운 음성이 추가되더라도 추가되는 기억장소가 필요하지 않는 숫자와 단모음 음성인식 시스템에 관하여 연구하였다.

그리고 인식률은 기존 표본 데이터로부터 개선된 표본 데이터 방식을 이용함으로써 기존 표본 데이터 방식의 신경회로망보다 높은 인식률을 얻을 수 있었다.

Ⅱ. 실시간 음성 데이터 저장법

음성신호는 피크치가 여기된 후 성도의 전달함수에 따라 변화하는 응답곡선으로, 음성인식을 위하여 정확한 시작점과 끝점을 실시간으로 검출하는것이 분석시간 단축과 인식률 향상에 중요하다[5].

지금까지 음성인식의 전처리 과정인 시작점과 끝점 검출법은 주로 ZCR, ENERGY 등의 방식을 이용하므로 실시간 저장이 곤란하다. 이러한 점 때문에 본 논문에서 시작점과 끝점 검출은 기억 영역을 3개 (START, DATA, END BUFFER) 구간으로 정하여 잡음 상태에서 임계치 조건을 체크하여 시작점과 끝점을 검출한다.

만약 임계치 조건을 만족하지 못하는 경우 시작점은 START BUFFER를 계속하여 임계치 조건을 체크하며, 조건을 만족하는 경우 그 때부터 ADDRESS 벡터를 이동시켜 DATA BUFFER 에 저장한다.

연속음은 음과 음사이의 휴지시간을 카운터하여 ADDRESS 벡터를 역으로 이동하여 불필요한 데이터를 제거한다.

따라서 실시간 시작점과 끝점 검출은 S/W의 간단한 조작으로 필요한 음성 데이터만으로 검출하였다. 이 경우 실시간 음성분석만 고려한다면 신경회로망을 이용한 실시간 음성인식 처리가 실현 가능하다.

Ⅲ. 역전파 학습알고리즘

역전파 학습은 통제 학습방식으로 학습시켜야 할 모든 입력에 대하여 각각의 원하는 출력과 실제 출력과의 차이를 최소화 하기위하여 회로망 가중치를 최적화하는 방식이다.

각 노드는 자신이 속하는 노드 보다 낮은 계층의 노드 출력을 입력으로 하며, SIGMOID 활성함수를 사용하여 입력값을 활성화시켜 출력값을 계산하고 자신보다 높은 계층의 노드에 전달한다. 즉 입력값이 입력되면 계산된 낮은 계층의 활성화 값을 바탕으로 상위계층의 출력이 차례로 계산되어 진행되며 최종적으로 출력층 노드의 출력값이 결정된다.

여기서 최적 가중치를 구하는 방법을 학습이라 하며 전향단계와 후향단계로 구분된다.

역전파 학습알고리즘은 다층 인식자(multi-layer perceptron)를 훈련하기위하여 사용하며, 다층 인식자의 바람직한 목표값과 실제 출력 활성값 사이에 평균 자승 에러를 최소로 하기 위한 알고리즘이다.

전향 단계는 신경회로망 입력 데이터를 입력하여 각 노드에 대하여 신경회로망 입력함수와 활성함수를 이용하여 출력을 산출하는 방식이다.

후향 단계는 원하는 출력과 실제 출력과의 차이를 계산하여 역전파 시키면 층과 층사이의 가중치가 조절되는 방식이다[6 7].

역전파 학습알고리즘은 다층 인식자에 DELTA 규칙을 일반화 한 것으로 다층 인식자 구조는 그림 1과 같고 역전파 학습 알고리즘은 다음과 같다.

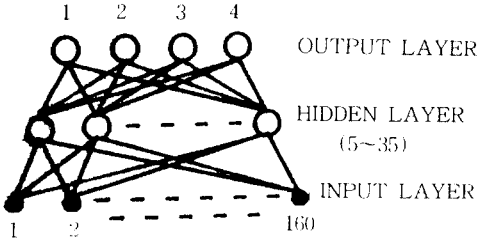


그림 1. 다층 인식자의 구조
The structure of MLP(multi layer perceptron)

알고리즘

제1단계 :

모든 노드의 연결강도 Wij와 각 노드의 OFFSETS 를 랜덤함수로 초기화 (W_{IJ} : 상위계층 J와 하위계층 i의 연결강도)

제2단계 :

입력 벡터와 출력 벡터를 제시한다.

제3단계 :

각 노드에 대하여 비선형 SIGMOID함수에 따라 출력 Oi를 계산한다.

$$O_I=f(net_I)=\frac{1}{1+\exp(-net_I)} \quad (1)$$

O_I : i층의 활성 출력함수

net_I : i층의 입력

제4단계 :

출력 노드의 오차를 계산하여 상위층의 노드j와 하위층 노드 i 사이 연결 가중치 W_{IJ}와 $bias_I$를 조절.

-조절규칙

· $W_{IJ}(n+1)=W_{IJ}(n)+\triangle W_{IJ}(n)$ (2)

$bias_I(n+1)=bias_I(n)+\triangle bias_I(n)$

n : time sequence

$\triangle W_{IJ}$: 상위층 노드 j와 하위층 노드 i의 연결 가중치 변화량

· $\triangle bias_I$: 상위층 노드 j와 하위층 노드 i의 바이어스 변화량

제5단계 :

오류 E가 기준치 보다 크면 제2단계로 가서 제5단계 까지의 과정을 반복, 적으면 가중치를 저장.

Ⅳ. 특징 축출

본 음성시스템에서 음성신호는 우선 저역필터(6차 체브체프)를 통과한 후 10KHZ 샘플링된 12 BIT A/D 변환기에 의하여 시작점과 끝점을 찾아낸다. 시작점과 끝점 검출 후 저장 데이터는 각 음성에 대하여 10 FRAME씩 저장된다.

저장된 음성 데이터는 한 FRAME(26.6ms) 당 프리엠파시스(0.98) 후 LPC방식 (16차)에 의하여 128개 해밍 윈도우 스펙트럼 파라메타를 추출하며, 총 10 FRAME 데이터(10 frame*128=1280)에서 1280개 스펙트럼 파라메타가 구하여진다.

추출된 스펙트럼 파라메타에 분산계수를 구하여 식 3과 같이 새로운 표본 데이터를 구하면 적은 표본 데이터에서 실제 많은 데이터를 표본한 것과 동일한 효과를 얻을 수 있다. 새로운 표본 데이터는 기존 표본 데이터의 T번째 FRAME 에서 I번째 스펙트럼값에 분산값을 첨가하여 만들어진다.

T번째 FRAME의 I번째 스펙트럼에 대하여 변형된 표본값 X_{It}는

$$X'it=Xit(1+Dit) \quad (3)$$

$Dit=RVit$

R=0.5, 0.5의 랜덤값

Vit=각 음의 T frame에서 i번째 LPC 스펙트럼 분산값

가 된다.

식 3에서 구한 X'it의 새로운 표본 데이터값은 역전파 신경회로망에 입력하기 위하여 10 FRAME 데이터의 1280개 LPC 스펙트럼값을 정하여진 주파수 간격에 의하여 한 FRAME 당 16CH로 나누어 모두 160개 데이터를 생성한다.

LPC 스펙트럼값에 대하여 최대값과 최소값을 구한 후, 신경회로망의 입력측에 입력하기 위하여 식 4와 같이 0과 1 사이의 값으로 정규화 한다[8 9 10 11].

이 때 사용되는 정규화 공식은

$$X=(D-MI)/(MA-MI) \quad (4)$$

D : LPC스펙트럼 데이터값
MA : 각 단음의 LPC스펙트럼 최대값
MI : 각 단음의 LPC스펙트럼 최소값

이다.

V. 음성 DATA BASE

본 연구에 사용한 음성 데이터는 숫자 0~9까지와 단모음 3자(아, 우, 에)로써 13자 단음을 4사람이 한 음성에 대하여 5번 반복하여 260개 음성 데이터를 인식 실험하였다.

신경회로망에서 각 음성 입력값에 대하여 원하는 출력값은 13개 이므로 각 음성에 CODE값을 부여하였다.

음성 DATA BASE 중 130개는 학습에 사용한 데이터(DB1)이며 나머지 130개는 학습하지 않는 데이터(DB2)이다.

VI. 실험결과 및 고찰

실험에 사용한 시스템 구성도는 그림 2와 같다.

음성신호는 4절에서 제시된 방법에 의하여 추출하며, 추출된 LPC 스펙트럼 파라메타는

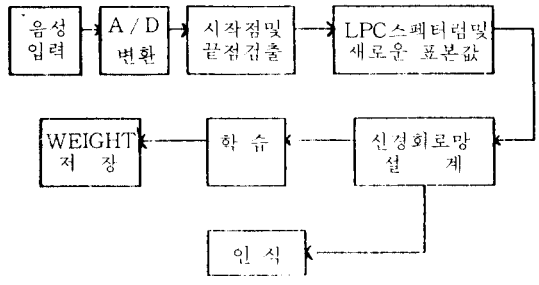


그림 2. 시스템 구성도
The system organization.

식 4에 의하여 정규화 된다.

신경회로망의 구성은 입력 160개, 출력 4개, 은익층 노드 25개로 설계하였다. 역전파 학습 알고리즘에서 학습률은 0.1, 관성항은 0.9로 학습을 수행하며, 가중치 Wit의 초기치는 −0.5, 0.5 사이의 랜덤값으로 학습하여 가중치를 결정하였다.

다층 인식자에서 DOMAIN KNOWLEDGE를 저장하는 은익층은 각 은익층에서의 노드의 수에 따라 신경회로망 성능에 많은 영향을 준다.

본 논문에서는 음성 인식시스템이 은익 노드 갯수의 변화에 따라 어떠한 인식과 에러 특성을 나타내는지 알아보기 위하여, 은익층의 노드 갯수를 5, 15, 25, 35개로 변화할 경우 각 음성의 학습 속도와 인식률에 대하여 분석하였다.

1) 은익층의 노드 갯수와 속도의 관계

일반적으로 노드의 갯수가 많을수록 계산량은 늘어나지만, 본 실험에서는 평가 단위를 계산시간이 아닌 LEARNING SWEEP (신경회로망에 1번 출현 후 1씩 증가 단위)으로 시스템이 안정될 때까지 몇 번의 LEARNING SWEEP가 일어나는 가를 비교하였다. 이 실험 결과, 그림 3에서와 같이 은익 노드의 갯수가 많을수록 LEARNING SWEEP는 감소하는 것으로 나타났다.

그림 4에서 은익 노드갯수에 대하여 LEARNING SWEEP 증가에 따라 신경회로망에서의 계산된 TSS(Total error sum of square)는 학습

이 진행할수록 에러가 감소된다.

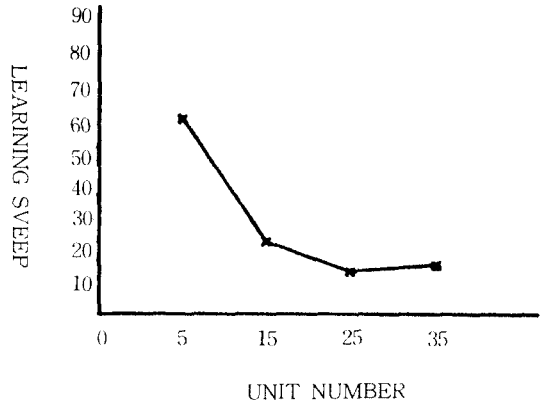


그림 3. 은익층의 노드갯수와 learning sweep 관계
Hidden layer node number and learning gweep relation

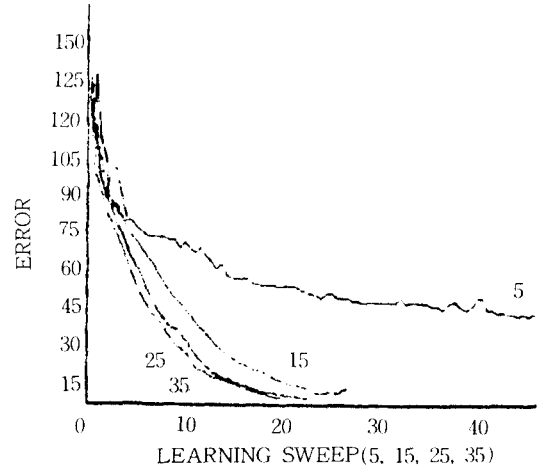


그림 4. 은익 노드의 수와 TSS 관계
The hidden node number and TSS relation

2) 은익층의 노드 갯수와 인식률의 관계

은익 노드의 갯수에 따라 같은 은익층에서도 입력값에 최적 노드가 구성될 경우 가장 높은 인식률을 얻을 수 있음을 본 논문에서 연구하였다. 은익 노드 갯수가 많으면 학습 속도가 빨라지는 현상은 노드들이 domain knowledge를 잘 표현할 수 있기 때문이라 생각된다.

은익 노드가 5개 일때 입력값에 대하여 domain knowledge를 표현할 수 없기 때문에, 학습하여도 인식률과 에러률은 개선되지 않는 local minima 현상이 발생되는 것이 분석되었다.

그림 5에서 실선은 학습된 데이터의 인식률이고, 점선은 학습되지 않은 인식률이다.

그림 6은 은익층 노드가 15, 25개 일 때 단음 "아"에 대한 인식률을 시뮬레이션이며, 은익층 노드가 15개 일 때 보다 25개 일 때가 양호한 인식률을 나타내고 있다.

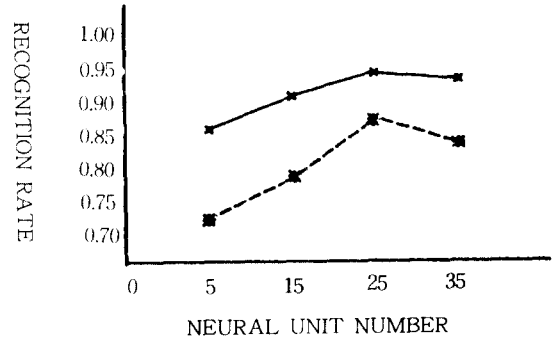
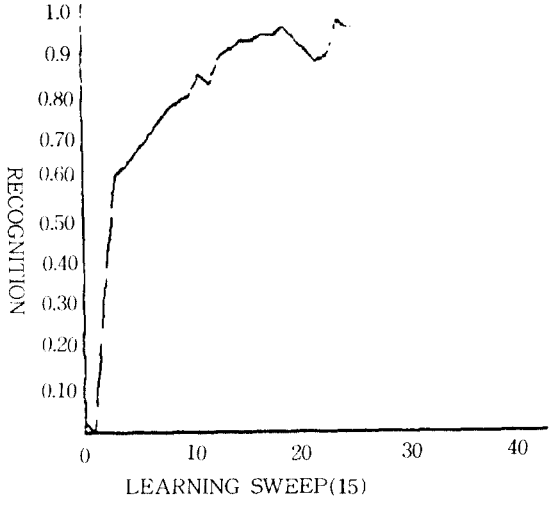


그림 5. 은익층의 노드 갯수와 인식률 관계
The hidden layer of the node number and recognition rate relation



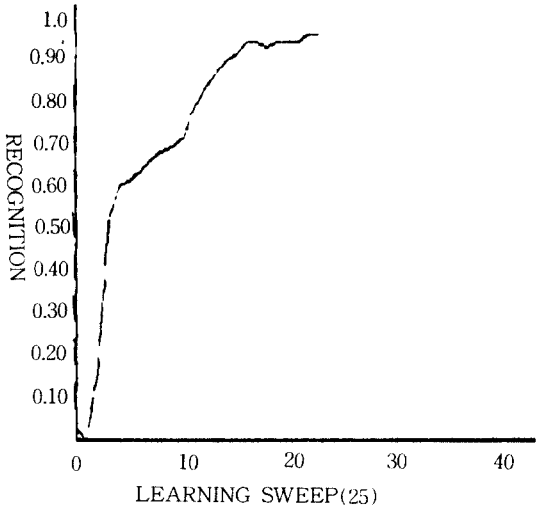


그림 6. 은익 노드 수에 따른 학습 진행상태의 인식률
The recognition rate of a learning state by the hidden node number

Ⅶ. 결 론

표 1은 은익 노드에 따라 학습된 데이터 인식률, 학습되지 않는 데이터의 인식률, 개선된 음성 학습 데이터 인식률을 평균값으로 구하였다.

표 2는 25개 은익 노드에서 각 표본 데이터의 인식률을 표현한 것으로 은익 노드 25개 일 때 학습속도, 인식률, 에러률이 가장 양호한 결과를 얻었으며, 학습된 데이터는 약 95%, 비 학습 데이터는 88%의 인식률을 나타내었다.

개선된 표본 방법에서 약간의 분산으로 적은 표본 데이터로 많은 표본 데이터가 표본된 결과를 얻을 수 있었다. 이러한 표본 방법은 음성 데이터에 약간의 잡음을 첨가하는 것이 실제적으

표 1. 실험 결과

은 익 노 드 수	learning sweep 수	학습음성 인식률	미학습음성 인식률	개선표본음성 학습인식률
5	60	85.5%	68.3%	88.3%
15	20	85.3%	79.2%	89.4%
25	15	93.6%	88.5%	95.1%
35	14	92.5%	88.3%	93.2%

표 2. 음성 인식률

인 식 률 데 이 터	비학습음성 인식률	학습음성 인식률	개선표본음성 인식률
1	86.4%	92.6%	96.5%
2	87.3%	93.4%	95.2%
3	89.2%	92.5%	94.0%
4	86.7%	94.4%	96.4%
5	88.0%	92.7%	94.3%
6	90.7%	96.5%	97.3%
7	88.4%	92.3%	93.7%
8	86.9%	92.4%	93.2%
9	89.4%	93.6%	95.6%
0	87.6%	92.8%	94.5%
아	90.4%	96.4%	96.9%
우	89.6%	93.8%	96.2%
에	87.3%	92.5%	94.7%

로 많은 데이터가 표본된 것처럼 되었다.

본 실험에서는, IBM 386(28 MHZ) 컴퓨터를 사용하였으나 학습 시간이 많이 걸리므로, 병렬 처리가 가능한 컴퓨터 및 반도체 칩을 사용한다면 실시간 처리가 가능하여 실용화 할 수 있을 것이다.

앞으로 문제점은 대상어의 확대로 학습 훈련의 시간 단축과 학습 DATA SETS 수의 증가에 대한 신경회로망의 확장성의 관계, 새로운 어휘에 관한 음성의 동적특성을 표현하는 방법등을 계속하여 연구해야할 점이다.

參 考 文 獻

1. R. P. Lippmann, "An introduction to computing with Neural Nets", IEEE ASSP Magazine, pp. 4~22, April, 1987.
2. D. E. Rumelhart, and J. L. Mcclelland, Parallel dist ributed processing : Explorations in the Microstructure of cognitron, VOL. 1 : Foundation, MIT Press / Bradford books, 1986.
3. N. H. Goddard, "Computing with structured neural network", IEEE computer, pp, 91~103, March, 198 8.
4. S. Y. Kumg and J. N. Hwang, "An algebraic proje-ction for optimal hidden unit size and learning rate in back propagation learning", IEEE International conference on neural network, Jul, 1988.
5. L. R. Rabiner "A linear predictive front-end processor for speech recognition in noise envrionments", Proc. ICASSP, pp. 1,324~1,327, 1987.
6. DE. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning representation by back-propagation errors", NUTURE, pp. 533~536, 1986.
7. J. J. Hopfield and D. N. Tnak, "Computing with neural circuits a model", Science, vol. 233, pp. 625~6 33, August, 1986.
8. T. Kohonen, "The nerual phonetic type writep", IEEE Computer vol. 21, No. 3, pp. 11~22, Mar. 1988.
9. J. Burr, "Experiments on Neural Net Recognition of spoken and written text", IEEE Tran. ASSP, vol. 36, No. 7, pp. 1,162~1,168, July, 1988.
10. J. L. Elman and D. Zipser, "Learning the hidden structure of speech", J. Acoust. SOC. AM 83.4, pp. 1,615~1,626, 1988.
11. Matsucka, "Syllable Recognition using intergrated neural net works", IEICE, vol. SP87~101, DEC. 1 989.

金 重 泰(Jung Tae KIM) 正會員

1957年10月13日生

1984年 2月 : 서울産業大學校 電子工學科 卒業(工學士)

1986年 2月 : 崇實大學校 大學院(工學碩士)

1986年 3月~現在 : 崇實大學校 大學院(博士課程修了)

1986年 3月~現在 : 富川工業專門大學電子通信科 專任講師