다층회귀신경망을 이용한 음성인식

어태경*, 배송학*, 김주성**, 안점영*
* 동의대학교 전자공학과
** 동아대학교 전자공학과

Speech Recognition Using Multilayered Recurrent Neural Networks

Tae-Kyung A*, Song-Hak Bae*, Joo-Sung Kim**, Jeom-Young Ahn*
* Dept. of Electronic Eng., Dong-eui Univ.
** Dept. of Electronic Eng., Dong-A Univ.

요 약

신경망에 의한 음절과 연속음성 인식시 동특성처리의 한방법으로 회귀신경망을 이용한다. 본 연구는 비회귀형 상위은닉층과 회귀형 하위은닉층을 가진 4층 구조의 다층회귀신경망(MLRNN)으로 예측기를 만들어 남성화자 5명이 CV형 음절 14개, CVC형 음절 14개를 각각 5회씩 발음한 총 700개의 음성중 3회분인 420개 음성으로 학습한 후 나머지 2회분인 280개 음성으로 인식을 평가한다.

입력신호의 예측차수와 상, 하위은닉층의 뉴런수를 변경시키면서 각각의 인식율을 조사해 본 결과 상위은닉층의 뉴런이 10개이고 하위은닉층의 뉴런이 10개와 15개 그리고 예측차수가 3, 4차일 때 가장 양호한 인식기로 동작한다는 것을 알 수 있었다. 이때 나타난 인식율은 Elman망 보다 다소 우세하다.

I. 서 론

신경망은 인간두뇌의 생리학적 구조와 기능을 모사한 인지적 정보처리장치로서 학습, 패턴인식, 연상기억능력이 있다. 사람의 두뇌는 고차원적인 비선형 시스템으로 작동하므로 이와 같은 두뇌기능을 가진 신경망으로 발전하려면 복잡한 동적기능을 수행할 수 있어야 한다. 신경망을 이용하여 동적인 정보를 처리하는 방법은 크게 두가지로 분류된다.

첫째, 시간정보를 공간정보로 변환한 후 처리하는 방법인데, TDNN(Time Delay Neural Network)[1], NETtalk[2]등이 여기에 속한다. 이 방법은 시간정보를 공간정보로 변환하는 전처리 작업, 적절한 프레임길이 결정, 신경망크기의 대규모화등의 문제점으로 시간정보 처리에 한계가 있다.

둘째, 회귀연결을 통해 시간정보를 그대로 처리하는 방법으로, 이때 회귀연결은 과거의 뉴런활성화(neuarl activation)과정을 기억하는 메모리의 역할을 한다. 이의 대표적인 것으로서 홉필드 모델을 시간정보의 연상에 응용한 TAM(Temporal Associative Memories)[3]과 다층퍼셉트론(Multilayered Perceptron : MLP)을 변형한 회귀신경망(Recurrent Neural Network : RNN)이 있다.

음성인식은 음소, 음절, 연속음성단위로 처리할 수 있는데, 음소인식은 종전의 전향신경망(Feedforward Neural Network)으로 처리 가능하지만, 음절이나 연속음성인식은 동적특성을 흡수하기 위하여 주로 RNN 모델을 이용한다[4][5].

본 연구에서는 음성인식을 위하여 비회귀형 상위은닉층과 회귀형 하위은닉층을 가진 4층구조의 다층회귀신경망(Multilayered Recurrent Neural Network : MLRNN)을 구성한다. 입력데이터는 1 프레임당 LPC melcepstrum 10차를 사용하고, 오차역전파 알고리듬

을 이용하여 MLRNN을 예측기로 학습 한 후 인식실험을 수행한다. 예측차수를 프레임 단위로 2차에서 4차까지 변경하고 동시에 상, 하위 은닉층의 뉴런수를 각각 5, 10, 15개로 변경하면서 CV형 14개 음절과 CVC형 14개 음절 그리고 이들을 합한 28개의 음절에 대한 음성인식율을 조사하여 2개의 은닉층을 가진 MLRNN의 인식기능을 알아보고자 한다.

II. 회귀 신경망

2.1 Jordan망과 Elman망

신경망 모델은 공통적으로 한 개의 뉴런이 다른 뉴런들과 연결가중치를 통해 상호연결되는 구조를 취하고 있지만 뉴런의 특성, 뉴런들의 연결형태를 나타내는 망 토폴로지, 그리고 연결가중치를 조절하는 학습규칙에 따라 모델의 특성이 달라진다[6]. 그 중에서 시간성보처리를 위한 신경망 모델들은 모두 회귀연결을 하고 있는 것이 특징이다. 그 대표적인 것으로서 Jordan망[7]과 Elman망[8]을 들 수 있다.

Jordan망은 그림 1과 같이 전형적인 MLP의 출력뉴런에서 상태뉴런으로 회귀연결을 하여 출력을 귀환하고 각 상태뉴런은 1보다 작은 고정된 계기의 자기회귀루프를 갖는 구조로 되어있다. 상태뉴런은 과거의 모든 상태입력에 대해 지수함수적으로 감소하는 합을 계산하는 일종의 기억소자로 작용한다.

Elman망은 그림 2와 같이 매 시간마다 은닉뉴런상의 활성패턴이 문맥층(context layer)으로 복사되어 다음 시간단계에서 회로망의 일부로 작용하는 구조를 이루고 있다. Elman망은 내부층에 기억되는 정보를 축적함으로써 입력의 문맥정보를 잘 반영할 수 있는 특징이 있다.

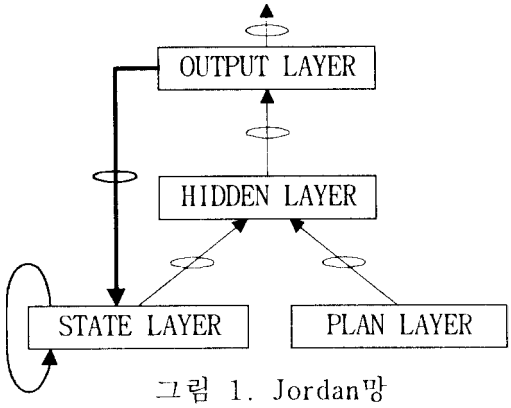

그림 1. Jordan망

Fig. 1. Jordan's recurrent network

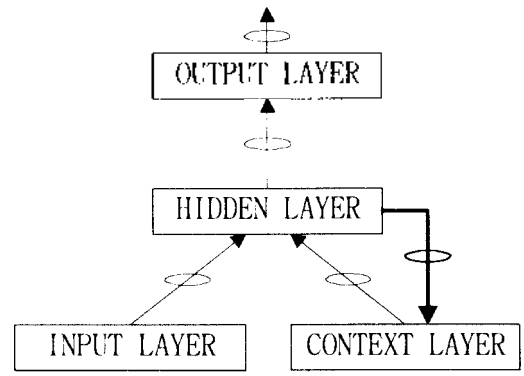

그림 2. Elman망

Fig. 2. Elman's recurrent network

오류역전파 알고리듬으로 Elman망을 학습하는 과정은 다음과 같다. 오류역전파 알고리듬은 처음 비회귀구조를 가진 다층전향신경망의 학습알고리듬으로 개발되었으나 회귀망에도 이를 적용 할 수 있다. 그러나 회귀망에서 역전파과정이 적절하게 진행되기 위해서는 회귀연결된 뉴런들의 그 이전시간에 대한 활성상태가 보존되어야 한다.

$S^T = \{s(1), \cdots, s(t), \cdots, s(T)\}$를 길이가 T프레임인 학습신호라 할 때, 한 프레임의 신호 $s(t) = [s_1(t), \cdots, s_N(t)]$는 시간 t에서의 N차원의 음성특징벡터이다. $s(t)$를 교사신호로 설정하고 그 이전시간의 신호 $s(t-1)s(t-2)\cdots$를 순차적으로 연결한 신호를 입력으로 인가하여 신호를 예측한다. 만약 예측차수가 t이면 입력신호는 $X(t) = [x_1(t), \cdots, x_{N_i}(t)]$인 N_i차원의 특징벡터가 된다.

은닉층의 뉴런수를 m개라 하고, 이 뉴런들의 출력을 모두 문맥층으로 복사하여 이를 다시 은닉층으로 귀환할 때 은닉층의 j번째 뉴런의 입력성분 $n_j(t)$와 출력성분 $h_j(t)$는 각각 다음과 같다.

$$n_j(t) = \sum_{i=0}^{m+N_i} w_{ji} z_i(t) \qquad j = 1, \cdots, m$$
$$h_j(t) = f(n_j(t)) \tag{1}$$

위 식에서 w_{ji}는 입력층의 i번째 뉴런에서 은닉층의 j번째 뉴런까지의 연결가중치이고, $f(\cdot)$는 비선형 시그모이드함수를 의미한다. $z_i(t)$는 시간 t에서의

입력벡터 X(t)와 복사된 은닉층의 출력성분벡터 H(t-1)을 연결한 벡터로 수식으로 표현하면 다음과 같다.

$$z_t(t) = [x_1(t), \cdots, x_{N_c}(t), h_1(t-1), \cdots, h_H(t-1)] \quad (2)$$

출력층에서 교사신호 $s(t)$에 대한 k번째 뉴런의 예측성분은 다음과 같다.

$$\hat{s}_k(t) = \sum_{j=1}^{m} w_{kj}(t) h_j(t) \quad k = 1, \cdots, N \quad (3)$$

여기서 w_{kj}는 은닉층과 출력층간의 연결가중치이고, 출력층에서는 선형함수를 사용하므로 출력층 뉴런의 실제출력성분 $o_k(t)$는 식(3)의 예측성분과 같다. 따라서 출력층 전체의 예측오차는 다음과 같다.

$$E(t) = \frac{1}{2} \sum_{k=1}^{N} [e_k(t)]^2 \quad (4)$$

여기서 $e_k(t) = s_k(t) - \hat{s}_k(t) = s_k(t) - o_k(t)$이다. 예측오차 $E(t)$를 다음식과 같이 연결가중치에 대해 미분하여 이 값으로 모든 연결가중치를 보정하는 것으로써 1회의 학습이 완료된다.

$$\triangle w_{kj}(t) = -\alpha \frac{\partial E(t)}{\partial w_{kj}} = \alpha e_k(t) o_k(t) h_j(t) \quad (5)$$

$$\triangle w_{ji}(t) = -\alpha \frac{\partial E(t)}{\partial w_{ji}} = \alpha \sum_{l=1}^{m} e_k(t) \frac{\partial h_l}{w_{ji}} \quad (6)$$

위 식에서 α는 학습계수로 상수이다. $E(t)$를 극소화하는 연결가중치를 구하기 위하여 경사강하(gradient descent)기법으로 학습을 계속한다.

2.2 다층회귀 신경망

본 연구를 위하여 그림 3과 같은 다층회귀신경망(MLRNN)을 구성한다. 하위은닉층의 정보를 입력측으로 귀환하여 문맥층을 형성함으로써 동특성을 지니게 한다.

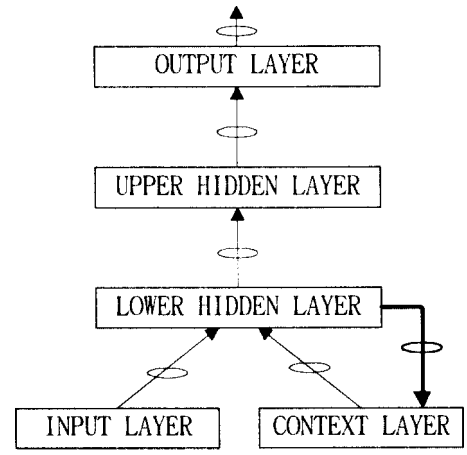

그림 3. 다층회귀신경망

Fig. 3. Multilayered Recurrent Neural Network

이 다층회귀신경망의 출력층과 상위은닉층은 선형출력함수, 하위은닉층은 비선형의 시그모이드함수를 사용하고, Elman망에서 이용한 오차역전파학습 알고리듬을 4층 MLP구조에 맞게 확장하여 적용한다.

III. 인식실험

3.1 음성자료 및 인식모델

MLRNN의 인식성능 실험대상은 한국어의 기본적인 CV형 14개음절과 CVC형 14개음절을 20대 남성화자 5명이 각각 5회씩 발성한 700개의 음성이다(표1). 이 음성중에서 3회분(420개)은 학습용으로 사용하고 나머지 2회분(280개)은 인식평가용으로 사용한다.

녹음된 음성은 12bit 양자화 레벨을 갖는 A/D변환기에서 10KHz로 샘플링된다. 이 신호를 H(z)=1-0.95 z^{-1}인 디지털필터로 고역강조한 후 크기가 20msec인 해밍창을 씌워 5msec씩 이동하면서 14차 LPC cepstrum 계수를 추출하고 이를 다시 10차 LPC melcepstrum 계수로 변환하여 본 연구의 특징파라미터로 사용한다.

표 1. 실험대상 한국어 음절

C V 형
가 나 다 라 마 바 사 아 자 차 카 타 파 하
C V C 형
갑 낙 단 란 만 반 산 안 잔 찬 칸 탄 판 한

예측차수에 따라 MLRNN의 입력층의 뉴런을 각각 20, 30, 40개로 하고, 상,하위은닉층의 뉴런을 각각

5, 10, 15개, 출력층의 뉴런을 10개로 한다.

MLRNN의 인식성능 수준을 알아보기 위하여 Elman망의 인식결과와 비교한다.

비교대상인 Elman망의 구성은 출력층 뉴런 10개, 은닉층의 뉴런 각각 5, 10, 15개 그리고 입력층의 뉴런은 예측차수에 따라 20, 30, 40개로 한다. 모든 학습은 3,000회까지 반복한다.

3.2 실험결과 및 고찰

예측차수와 상, 하위은닉층의 뉴런수를 변화시키면서 CV, CVC, CV+CVC음절에 대하여 인식실험을 하고 그 결과를 표2에 나타내었다.

학습된 각 MLRNN의 입력층에 인식대상 음성신호를 인가하여 출력층에 나타나는 평균예측오차를 각 망마다 구하고 그 값이 최소가 되는 망을 인식모델로 선정하여 인식율을 계산하였다.

이 결과에 의하면 신경망의 구조에 따라 인식성능에 상당한 차이가 있으며 이 MLRNN은 상위은닉층 뉴런이 10개, 하위은닉층의 뉴런이 10개와 15개 그리고 예측차수가 3차, 4차일 때 가장 양호한 인식기로 동작하고, 이때 최대 인식율은 CV음절에서 87.86%, CVC음절에서 84.29% 그리고 CV+CVC음절에서 80.71%이다.

전체적인 인식경향은 CV>CVC>CV+CVC 순으로 비교되는데, CVC는 음향학적으로 CV보다 동특징을 많이 포함하고, CV+CVC는 인식을 위한 비교대상수가 CV 또는 CVC보다 많기때문에 나타나는 결과이다. 부분적으로는 CV음절인 경우 상위은닉층 뉴런 5개, 하위은닉층 뉴런 10개 일 때 인식율이 향상되고, CVC음절인 경우는 상위은닉층 뉴런 15개, 하위은닉층 뉴런 15개일 때 비교적 양호한 인식성능을 나타내었다.

표3은 Elman망에 대한 인식실험 결과이다. 이 실험은 MLRNN의 인식성능과 비교하기 위하여 실시한 것이다. 두 망은 구조적인 차이 때문에 직접적인 비교는 불가능하지만, MLRNN을 실용적인 음성인식기로 이용한다고 가정하면 인식성능이 가장 양호한 구조를 택할 것이다. 그러므로 여기서도 인식성능이 양호한 상위은닉층 뉴런 10개 즉 표2의 중간 표시부분과 표3의 데이터를 비교하였다. 비교대상 인식율은 모두 27개이다. Elman망 보다 우세한 경우가 15, 열세한 경우가 11 그리고 동등한 경우가 1이므로 전체적으로 볼 때 MLRNN의 인식성능이 Elman망보다 우세하다는 것을 알 수 있다. 특히 상위은닉층 뉴런 10개, 하위은닉층 15개, 예측차수 4차인 경우 MLRNN의 인식율이 Elman망보다 5.27% 정도 상승한다.

표 2. MLRNN의 인식율

상위은닉층 뉴런수	하위은닉층 뉴런수	예측차수	인식율 (%) CV	CVC	CV+CVC
5	5	2	79.29	72.86	70.00
		3	76.43	67.86	66.79
		4	65.71	75.71	67.86
	10	2	83.57	74.24	73.93
		3	85.00	72.86	78.21
		4	77.86	77.14	77.14
	15	2	80.00	73.57	72.50
		3	84.29	75.71	76.79
		4	76.43	74.29	75.36
10	5	2	85.00	76.43	75.00
		3	76.43	72.14	68.93
		4	77.14	75.71	72.14
	10	2	82.86	69.29	72.86
		3	87.14	84.29	80.71
		4	87.86	82.14	79.29
	15	2	73.57	69.29	67.50
		3	86.43	84.29	80.36
		4	87.86	82.14	80.71
15	5	2	83.57	75.00	74.64
		3	77.14	72.86	67.86
		4	74.29	75.00	71.43
	10	2	78.57	72.14	70.36
		3	83.57	70.71	71.79
		4	75.71	73.57	71.79
	15	2	77.14	81.43	73.21
		3	75.71	82.86	71.79
		4	72.14	87.14	77.50

표 3. Elman망의 인식율

은닉층 뉴런수	예측차수	인식율 (%) CV	CVC	CV+CVC
5	2	77.14	73.57	67.86
	3	82.14	71.43	70.00
	4	81.42	76.43	71.07
10	2	84.28	80.00	76.43
	3	86.42	80.00	80.00
	4	83.57	85.00	79.64
15	2	79.29	72.14	71.07
	3	82.85	75.71	72.86
	4	83.57	82.14	78.93

Ⅳ. 결 론

본 연구는 비회귀형 상위은닉층과 회귀형 하위은닉층을 가진 4층구조의 다층회귀신경망(MLRNN)을 구성하여 음성인식 실험을 하고 그 인식결과로부터 한국어 단음절 음성인식에 적합한 MLRNN의 구조를 찾아서 이를 Elman망의 인식성능과 비교하였다.

실험결과 MLRNN 역시 다른 신경망처럼 구조와 대상음성에 따라 인식성능에 상당한 차이가 있으며

특히 상위은닉층 뉴런이 10개, 하위은닉층의 뉴런이 10개와 15개 그리고 예측차수가 3, 4차일 때 대체적으로 양호한 인식가로 동작하였다. 상위은닉층 뉴런이 10개일 때 MLRNN의 인식성능이 Mlman망에 비해 다소 향상되었다.

연구를 시작하기 전 예상했던 수준에 미달하였으므로 앞으로 알고리듬이나 망 토포리지능을 계속 연구하여 성능을 더욱 개선할 계획이다.

참 고 문 헌

1, A. Waibel, T.Hanazawa, G. Hinton, K. Shikano, K. J. Lang, "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans. on ASSP., Vol. 37, pp. 328-339, March 1989

2. T. J. Sejnowski, C. R. Rosenberg, "Parallel Networks that Learn to Pronounce English Text", Complex syst., Vol. 1, pp. 145-168, 1987

3. Kosko, Bart, Neural Networks and Fuzzy Systems, Prentice-Hall International Inc., 1992

4. R. J. Williams, D. Zipser, "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks", Neural Computation, pp. 270-280, 1989

5. 유제관, 나경만, 임재열, 안수길, "회귀신경예측 모델을 이용한 음성인식", 전자공학회 하계종합학술대회 논문집, 제18권 1호, 1995

6. R. P. Lippmen, "An introduction to Computing with Neural Nets", IEEE ASSP Magazine, pp. 4-22, Apr. 1987

7. M. I. Jordan, "Serial Order : A Parallel Distributed Processing Approach", Technical Report ICS-8604, Institute for Cognitive Science, University of California, San Diego, La Jolla, California, May 1986

8. J. L. Elman, "Finding Structure in Time", Technical Report CRL-8801, Center for Research in Languge, University of California, San Diego, La Jolla, California, Apr. 1988