1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

다층회귀신경망의 회귀구조에 따른 음성인식성능 비교

Comparison of the Speech Recognition Performance based upon the Recurrent Structure of the Multilayered Recurrent Neural Network

이 태경*, 배 송학*, 안 점영*

(Tae Kyung A*, Song-Hak Bae*, Jeom Young Ahn*)

* 동의대학교 전자공학과

요 약

4층구조인 다층퍼셉트론으로 부터 입력층을 제외한 각 층의 출력성분을 하위은닉층으로 귀환하는 3모델의 다층회귀신경망을 구성하고, 각 모델별 망의 크기에 따른 음성인식성능을 분석 비교한다. 과거의 입력신호를 출력층에서 예측하여 오차신호를 계산하고, 이 오차신호가 최소화하는 방향으로 연결세기를 조정한다. 실험결과 3회귀모델중 상위은닉층의 회귀연결방식이 가장 양호한 인식율을 나타내었으며, 각 망 공히 상, 하위은닉층의 뉴런수 10, 15개, 예측차수 3, 4차 일 때 인식성능이 양호하였다. 그리고 회귀신경망이 비회귀신경망에 비해 인식율이 크게 향상된다는 것을 확인 할 수 있었다.

I 서 론

신경망이 지닌 고도의 병렬성, 비선형적 특성 그리고 학습의 기능을 이용하면 패턴인식이 가능하다. 음성은 화자, 발성시의 분위기, 단독음과 연속음 기타 많은 요인에 따라 특성이 민감하게 변화하므로 이와 같은 동적특성을 효과적으로 처리하려면 전처리 과정에서 음성데이터를 변화시키던지 아니면 망의 구조를 변경해야 한다.

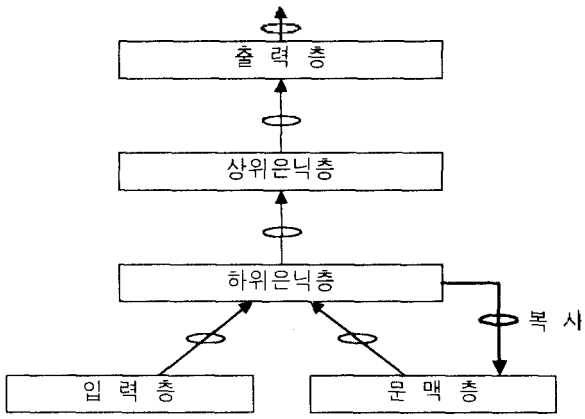
동적정보처리를 위하여 TDNN(Time-Delay Neural Network)[1], ATNN(Adaptive TDNN)[2]은 전향신경망에 시간지연의 개념을 도입하였으며, Jordan망[3]과 Elman망[4]은 전향신경망에 귀환연결을 부가하여 과거의 이력이 현재의 출력에 반영되도록 하였다.

본 연구에서는 4층구조의 기본 전향신경망으로 부터 하위은닉층의 출력을 다시 하위은닉층으로, 상위은닉층의 출력을 하위은닉층으로 마지막으로, 출력층의 출력을 하위은닉층으로 귀환하는 세종류의 다층회귀신경망(Multilayered Recurrent Neural Network : MLRNN)을 구성한다. CV형 음절 14개와 CVC형 음절 14개의 음성으로 구성된 각 망의 음성인식율을 조사하고 회귀구조에 따른 인식성능을 상호비교한다. 그리고 기본 전향신경망의 인식율과 비교하여 회귀형 신경망이 비회귀형에 비해 어느 정도 인식율이 향상되는가를 알아본다.

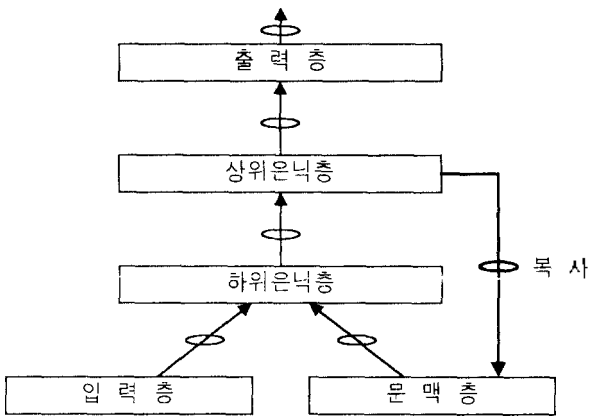
II. 다층회귀신경망

Jordan망은 3층구조로 된 다층퍼셉트론의 출력층의 출력을 복사하여 상태층(state layer)을 형성하고 상태층의 1차 지연된 신호를 다음 입력신호와 함께 은닉층으로 제시하는 구조를 취한다. 반면에 Elman망은 은닉층의 활성신호를 복사하여 문맥층(context layer)을 형성하고 문맥층의 1차 지연된 신호를 다음 입력신호와 함께 다시 은닉층으로 회귀하는 구조이다. 회귀연결을 통해 신호의 문맥정보를 잘 처리할 수 있으므로 두 망은 음성처럼 동특성이 많이 내포된 신호의 순차처리에 효과적이다.

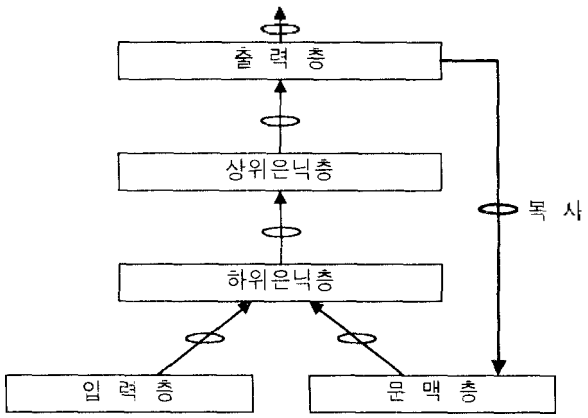
본 연구를 위하여 구성한 다층회귀신경회로망은 그림 1과 같다. 4층구조의 다층퍼셉트론에 회귀연결을 부가함으로써



(a) 하위은닉층의 출력을 귀환한 신경망



(b) 상위은닉층의 출력을 귀환한 신경망



(c) 출력층의 출력을 귀환한 신경망

그림 1. 다층회귀신경망

과거의 뉴런활성화(neural activation)과정을 기억하게 한 것이다. 하위은닉층은 시그모이드형 전달함수 그리고 상위은닉층과 출력층은 선형 전달함수를 사용하며 귀환성분은 문맥층에서 1차 지연되어 다음시간의 입력과 함께 하위은닉층으로 입력된다. 과거의 음성패턴으로부터 예측한 출력패턴을 현재의 패턴(목표출력패턴)과 비교하여 오차를 최소화하는 방향으로 학습한다.

III. 인식실험

3.1 실험 데이터

한국어의 기본적인 CV형 음절 14개와 CVC형 음절 14개[표1]을 20대 남성화자 5명이 각각 5회씩 발성한 700개의 음성중 3회분(420개)은 학습용으로 사용하고 나머지 2회분(280개)은 인식평가용으로 사용한다.

녹음 된 음성을 12bit 양자화 레벨을 갖는 A/D 변환기에서 10KHz로 샘플링된다. 이 신호를 $H(z)=1-0.95z^{-1}$인 디지털필터로 고역강조한 후 프레임 2msec인 Hamming창을 씌워 5msec씩 이동하면서 14차 LPC cepstrum계수를 추출하고 이를 다시 10차 LPC melcepstrum계수로 변환하여 실험데이터로 활용한다.

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
간 난 단 란 만 반 산 안 잔 찬 칸 탄 판 한

3.2 실험결과 및 고찰

그림 1의 다층회귀신경망의 출력층은 10차원의 벡터성분으로 표시되므로 출력층의 뉴런수는 10개로 고정하고 그 대신 상, 하위은닉층의 뉴런수가 인식율에 미치는 영향을 알아보기 위하여 상, 하위은닉층의 뉴런수를 각각 5, 10, 15개까지 변경한다. 입력층의 뉴런수는 예측차수에 따라 결정되며 예측차수당 10개씩 할당하고 차수를 2, 3, 4차까지 변경한다.

각 음성데이터는 제1 프레임에서, 제2, 제3 프레임순으로 예측차수만큼 입력패턴으로 지정하고 그 다음 프레임을 목표출력패턴으로 설정한다. 이 경우 첫 번째와 마지막 프레임 데이터를 예측차수 만큼 복사하여 모든 데이터가 입출력으로 입력과 목표출력 패턴의 역할을 하도록 한다.

모든 음성이 학습되면 전체적으로 음성개수만큼의 서로 다른 연결강도를 가진 신경망이 구성된다. 이들 망에 인식

하고자 하는 음성데이터를 입력하여 각 망의 출력층에 나타나는 평균예측오차를 계산하여 이 값이 최소가 되는 망을 인식망으로 선정하여 인식율을 계산한다.

그림 1의 구조를 가진 다층회귀신경망과 4층으로 구성된 다층퍼셉트론(비회귀형)의 인식실험결과는 표 2와 같다.

인식결과를 간략하게 비교분석하면 다음과 같다.

(1) 다층퍼셉트론

망의 구조에 따라 최저 36.07%에서 최대 85%까지 인식율의 분포범위가 넓다.

특정구조 즉 상위은닉층의 뉴런이 10개, 하위은닉층의 뉴런이 15개, 예측차수 4차에서 CV, CVC, CV+CVC 음절에 대해 각각 84.29%, 85%, 80.71%의 인식율을 나타내지만 평균적으로 각각 66.06%, 63.38%와 59.51% 정도의 수준이므로 본 연구대상의 음성인식기로 부적합하다.

(2) 하위은닉층의 출력을 하위은닉층으로 회귀한 경우

인식율의 변동폭이 다층퍼셉트론만큼 크지 않고, 평균인식율이 다층퍼셉트론 보다 CV음절에서 13.6%, CVC음절에서 12.67%, CV+CVC음절에서 14.06% 정도 상당량 상승한다. 따라서 회귀연결을 통해 문맥정보의 처리가 잘 이루어진다는 것을 알 수 있다.

(3) 상위은닉층의 출력을 하위은닉층으로 회귀한 경우

이 모델은 하위은닉층에서 회귀한 경우보다 평균인식율이 CV음절에서 3.49%, CVC음절에서 1.38%, CV+CVC음절에서 1.6%정도 향상되며, 세종류의 회귀모델중 가장 성능이 우수한 신경망이다. 하위은닉층의 뉴런이 10개, 상위은닉층의 뉴런이 10개, 15개인 때 성능이 가장 양호하다.

(4) 출력층의 출력을 하위은닉층으로 회귀한 경우

이 모델의 인식결과는 하위은닉층 회귀모델의 결과보다 향상되지만 상위은닉층 회귀모델에 비해 인식성능이 다소 떨어진다. 2개의 은닉층과 출력층을 거쳐 나감으로써 더욱 더 세분된 패턴으로 분류되어 오히려 인식율이 저하된다고 생각된다.

Ⅳ. 결 론

본 연구에서는 회귀신경망의 동특성처리 능력과 회귀구조에 따른 음성인식성능을 알아보기 위하여 4층의 다층퍼셉트론 구조에서 하위은닉층, 상위은닉층과 출력층의 출력을 각각 하위은닉층으로 귀환하는 3 모델의 다층회귀신경망을 구성하고, 각 망에 대한 음성인식 성능을 실험을 통해 비교분석해 보았다.

실험결과 회귀형 신경망은 비회귀형에 비해 동특성처리 능력이 우수하고, 구성한 3 모델의 회귀신경회로망중에서 상위은닉층의 출력성분을 하위은닉층으로 회귀할 때 인식성능이 가장 양호하고, 이와같은 다층회귀신경망으로 음절단위의 음성을 인식하는 경우 상, 하위은닉층의 뉴런이 10 혹은 15개, 예측차수가 3 혹은 4차일 때 비교적 양호한 인식률로 동작한다는 것을 알 수 있었다.

인식율 향상을 위하여 이 다층회귀신경망의 구조에 적합한 학습 알고리듬을 연구할 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] D. T. Lin, J.E. Dayhoff, and P. A. Ligomenides, "Adaptive time-delay neural network for temporal correlation and prediction," SPIE Intelligent Robots and Computer Vision XI: Biolocal, Neural Net, and 3-D Method, vol. 1826,(Boston, November), pp.170-181, 1992

[3] 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

[4] 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

[5] 유제관, 나경민, 임재열, 안수길, "회귀신경예측 모델을 이용한 음성인식", 전자공학회 하계종합학술대회 논문집, 제18권 1호, 1995

[6] 어태성, 배송학, 김주성, 안점영, "다층회귀신경망을 이용한 음성인식", 제15회 음성통신 및 신호처리 워크샵 논문집, KSSP'98 Vol 15. No. 1, pp. 267-268, 1998

표 2. 다층회귀신경망과 다층퍼셉트론의 음성인식율

상위층 뉴런수	하위층 뉴런수	예측 차수	다층퍼셉트론			하위은닉층에서 귀환			상위은닉층에서 귀환			출력층에서 귀환		
			CV	CVC	CV+CVC	CV	CVC	CV+CVC	CV	CVC	CV+CVC	C V	CVC	CV+CVC
5	5	2	53.57	57.14	49.29	79.29	72.86	70.00	77.86	69.29	68.93	75.00	75.71	68.57
		3	57.14	45.71	43.93	76.43	67.86	66.79	75.00	72.14	67.85	77.86	69.29	68.21
		4	47.86	40.71	36.07	65.71	75.71	67.86	80.00	75.00	71.07	80.00	72.86	71.07
	10	2	67.14	62.86	57.14	83.57	74.24	73.93	82.14	72.14	73.21	84.26	72.86	74.29
		3	68.57	76.42	66.43	85.00	72.86	78.21	85.00	73.57	75.00	82.14	75.71	74.64
		4	70.00	68.57	62.50	77.86	77.14	77.14	83.57	75.71	74.29	85.00	77.86	76.79
	15	2	82.86	71.43	72.50	80.00	73.57	72.50	83.57	75.00	74.64	83.57	73.57	75.00
		3	80.71	73.57	74.26	84.29	75.71	76.79	85.00	73.57	73.57	85.71	77.14	77.14
		4	79.29	74.29	72.86	76.43	74.29	75.36	86.43	77.86	77.86	86.43	80.00	79.29
10	5	2	50.71	43.57	38.93	85.00	76.43	75.00	82.86	72.86	72.86	85.00	71.43	73.57
		3	55.71	52.14	46.43	76.43	72.14	68.93	80.00	76.43	72.86	77.86	76.43	72.14
		4	48.57	37.14	36.79	77.14	75.71	72.14	80.00	76.43	72.86	81.43	76.43	74.64
	10	2	60.71	58.57	62.86	82.86	69.29	72.86	84.26	75.71	76.79	84.26	84.26	82.86
		3	72.14	69.28	76.71	87.14	84.29	80.71	86.43	78.57	77.14	82.86	80.00	76.79
		4	69.28	70.71	62.14	87.86	82.14	79.29	87.14	82.86	80.71	83.57	82.14	77.50
	15	2	82.86	85.00	80.35	73.57	72.86	67.50	85.00	86.43	82.14	86.43	81.43	84.26
		3	83.57	80.71	78.57	86.43	81.43	80.36	90.71	83.57	83.93	88.57	81.43	83.57
		4	84.29	85.00	80.71	87.86	84.29	80.71	85.00	88.57	81.76	88.57	76.43	73.93
15	5	2	47.86	40.00	36.07	83.57	75.00	74.64	84.29	80.00	72.50	79.29	71.43	80.00
		3	55.71	52.14	47.50	77.14	72.86	67.86	79.29	74.29	70.00	77.86	74.29	68.93
		4	51.43	41.43	37.14	74.29	75.00	71.43	82.14	77.86	75.00	81.43	75.00	72.86
	10	2	55.00	62.14	50.35	78.57	72.14	70.36	81.43	73.57	70.71	80.71	68.57	67.86
		3	63.57	67.86	63.21	83.57	70.71	71.79	86.43	80.71	77.14	87.14	73.57	75.71
		4	67.14	68.57	63.57	75.71	73.57	71.79	80.00	72.14	71.07	86.43	72.86	73.21
	15	2	80.00	79.29	74.64	77.14	81.43	73.21	82.14	76.43	75.00	78.57	79.29	72.86
		3	79.29	76.43	73.57	75.71	82.86	71.79	85.00	87.14	82.86	77.14	82.86	76.07
		4	68.57	70.71	62.50	72.14	87.14	77.50	84.26	82.86	77.86	82.86	83.57	77.86
평균인식율			66.05	63.38	59.51	79.65	76.05	73.57	83.14	77.43	75.17	82.59	76.53	75.17