다층회귀예측신경망의 음성인식성능에 관한 연구

안 점 영*

A Study on the Speech Recognition Performance of the Multilayered Recurrent Prediction Neural Network

Jeom-Young Ahn*

이 논문은 1998년도 동의대학교 학술연구조성비에 의해 연구되었음

요 약

4층구조의 다층퍼셉트론을 변형하여 3 종류의 다층회귀예측신경망을 구성하고, 예측차수, 두 은닉층의 뉴런 개수, 연결세기의 초기치 및 전달함수 변화에 따른 각 망의 음성인식성능을 실험을 통해 각각 비교분석한다.

실험결과에 의하면, 다층회귀신경망이 다층퍼셉트론에 비해 음성인식성능이 우수하다. 그리고 구조적으로는 상위은닉층의 출력을 하위은닉층으로 회귀할 때 인식성능이 가장 우수하며, 각 망 공히 상, 하위은닉층의 뉴런 10 혹은 15개, 예측차수 3 혹은 4차일 때 인식률이 양호하다. 학습시 연결세기의 초기치를 -0.5에서 0.5사이로 설정하고, 하위은닉층에서 단극성 시그모이드 전달함수를 사용할 때 인식성능이 더욱 향상된다.

Abstract

We devise the 3 models of Multilayered Recurrent Prediction Neural Network(MLRPNN), which are obtained by modifying the Multilayered Perceptron(MLP) with 4 layers. We experimentally study the speech recognition performance of 3 models by a comparative method, according to the variation of the prediction order, the number of neurons in two hidden layers, initial values of connecting weights and transfer function, respectively.

By the experiment, the recognition performance of each MLRPNN is better than that of MLP. At the model that returns the output of the upper hidden layer to the lower hidden layer, the recognition performance shows

* 동의대학교 전기 · 전자공학부

접수일자 : 1999년 2월 11일

the best value. All MLRPNNs, which have 10 or 15 neurons in the upper and lower hidden layer and is predicted by 3rd or 4th order, show the improved speech recognition rate. On learning, these MLRPNNs have a better recognition rate when we set the initial weights between -0.5 and 0.5, and use the unipolar sigmoid transfer function in the lower hidden layer.

I. 서 론

음성인식이란 시스템에 입력된 음성을 기계가 정확히 인식하여 문자로 바꾸어 주거나 혹은 입력된 음성을 이해하여 적절하게 대응해 주는 것을 말한다. 음성인식 수준은 음소와 음소 혹은 음절과 음절사이에 나타나는 조음현상, 발성시간차, 기타 음운적 특징에 따른 동적특성 부분의 처리기법에 좌우된다.

신경망을 이용하여 음성의 동적정보를 처리하는 기법은 두가지로 대별된다. 하나는, 시간지연요소를 사용하여 시간정보를 공간정보로 변환한 후 처리하는 방법으로 TDNN(Time-Delay Neural Network)[1], NETalk[2]등이 있다. 이 방법은, 시간정보를 공간정보로 변환하는 전처리과정이 필요하고, 적절한 프레임 길이를 결정해야 하며, 프레임 길이에 따른 신경망의 크기가 대규모화 되는 문제점이 있다.

다른 하나는, 회귀연결을 통해 시간정보를 직접 처리하는 방법으로 다층퍼셉트론(Multilayered Perceptron : MLP)을 변형한 회귀신경망(Recurrent Neural Network)[3][4]이 있다. 회귀신경망은 문맥정보를 나타낼 수 있는 내부상태를 가지므로 비선형성을 갖는 동적시스템을 모델링하는데 적합하며, 이때 회귀연결이 과거의 뉴런활성화(Neural activation)과정을 기억하는 메모리역할을 한다.

예측형 신경망[5][6]은 서로 다른 길이를 가진 음성을 시간축상에서 정규화할 필요가 없으므로 대단위 음절을 인식할 수 있으며, 회귀연결시 더 나은 예측성능을 기대할 수 있다.

본 연구는 4층구조의 다층퍼셉트론을 회귀신경망으로 구조변경할 때 음성인식수준이 어느 정도인가를 알아보기 위해 4층구조에서 하위은닉층의 출력을 다시 하위은닉층으로, 상위은닉층의 출력을 하위은닉층으로, 마지막으로 출력층의 출력을 하위은닉층으로 귀환하는 3 종류의 다층회귀신경망을 구성하고 한국어의 기본적인 CV형 음절 14개와 CVC형 음절 14개에 대한 인식실험을 수행한다.

그리고 이 다층회귀신경망은 입력층에 가해진 음성특징벡터를 출력층에서 예측하는 예측기로 활용하고, 예측차수, 망의 크기(두 은닉층의 뉴런수), 연결세기의 초기치 및 전달함수를 변화시키면서 그에 따른 인식성능을 조사 비교한다.

Ⅱ. 다층회귀예측신경망의 구성과 학습

2.1 신경망 구성

본 연구를 위하여 구성한 다층회귀예측신경망(Multilayered Recurrent Prediction Neural Nerwork : MLRPNN)은 그림1과 같다. 회귀연결을 통해 과거의 신호를 현재의 신호에 반영되도록 하였으며, 귀환된 성분을 문맥층(Context layer)이라 하고, 문맥층에서 1차 지연된 신호를 다음 입력신호와 함께 하위은닉층으로 입력한다.

2.2 학 습

기존의 오류역전파 알고리듬으로 학습한다. 오류역전파 학습 알고리듬은 처음 비회귀구조를 가진 다층전향신경망의 학습알고리듬으로 개발되었으나 회귀망에도 이를 적용할 수 있다. 회귀망에서는 전향연결만 학습하고 귀환연결은 학습하지 않는다.

하위은닉층은 시그모이드형 전달함수, 그리고 상위은닉층과 출력층은 선형전달함수를 사용한다. 시그모이드형 전달함수는 결정영역(Decision region)이 직선이 아닌 완만한 곡선이기 때문에 미분이 가능하여 신경망학습에 많이 이용된다.

단극성과 양극성 시그모이드함수는 각각 식(1), (2)와 같다.

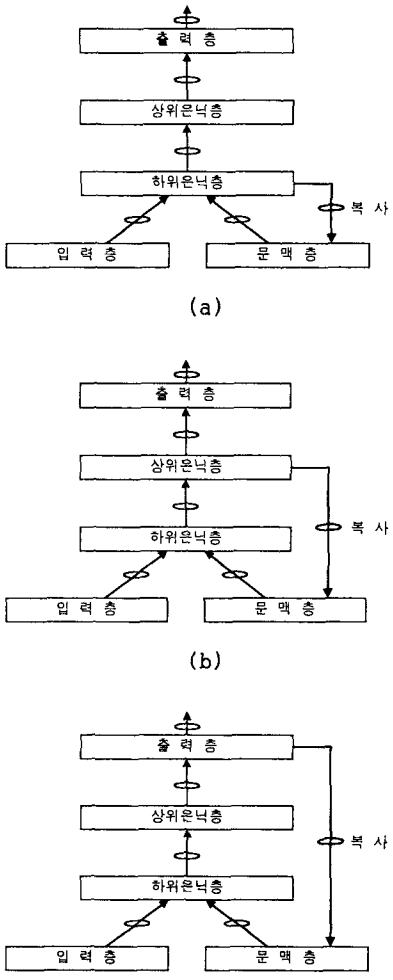
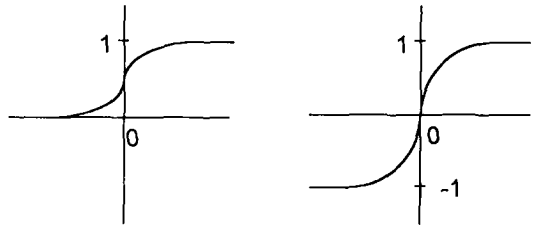


그림 1. 다층회귀예측신경망 (a) 하위은닉층 출력 회귀 신경망 (b) 상위은닉층 출력 회귀신경망 (c) 출력층 출력 회귀신경망

Fig. 1 Multilayered recurrent prediction neural network
(a) Neural network returning the output of the lower hidden layer
(b) Neural network returning the output of the upper hidden layer
(c) Neural network returning the output of the output layer



(a) 단극성 (b) 양극성

그림 2. 시그모이드형 전달함수

Fig. 2 Sigmoid type transfer function
(a) Unipolar (b) Bipolar

$$f(net)=\frac{1}{1+\exp(-\lambda net)} \quad \cdots\cdots (1)$$

$$f(net)=\frac{2}{1+\exp(-\lambda net)}-1 \quad \cdots\cdots (2)$$

여기서 $\lambda(>0)$는 net=0 근처에서 연속함수 $f(net)$의 첨예도를 결정하는 뉴런의 이득에 비례한다.

Ⅲ. 인식실험

3.1 음성 데이터

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

녹음된 음성은 12bit 양자화 레벨을 갖는 A/D 변환기에서 10KHz로 샘플링된다. 이 신호를 $H(z)=1-0.95z^{-1}$인 디지털필터로 고역강조한 후 크기가 20msec인 Hamming창을 씌워 5msec씩 이동하면서 14차 LPC cepstrum계수를 추출하고 이를 다시 10차 LPC melcepstrum계수로 변환하여 실험데이터로 활용한다.

표 1. 실험대상 한국어 음절

Table 1. Korean syllables for experiment

CV	가 나 다 라 마 바 사 아 자 차 카 타 파 하
CVC	간 난 단 란 만 반 산 안 잔 찬 칸 탄 판 한

3.2 실험결과 및 고찰

그림1의 다층회귀예측신경망은 출력층이 10차원의 백터성분이므로 출력층의 뉴런수는 10개로 고정하고 그 대신 상·하위은닉층의 뉴런수에 따라 인식률의 변화를 알아보기 위해 상·하위은닉층의 뉴런수를 각각 5, 10, 15개까지 변경한다. 문맥층의 뉴런수는 회귀되는 층의 뉴런수와 같고, 입력층의 뉴런수는 예측차수에 따라 결정되며 예측차수당 10개씩 할당한다. 예측차수의 변화에 따른 인식률 변화를 알아보기 위해 예측차수를 2, 3, 4차까지 변경시킨다.

연결세기의 초기치를 0에서 1.0까지 또는 -0.5에서 0.5까지 불규칙하게 주었을 때 그리고 하위은닉층의 전달함수가 그림2의 단극성과 양극성일 때 인식률 변화도 함께 알아본다.

학습이 종료되면 전체적으로 음성개수만큼의 서로 다른 연결세기를 가진 신경망이 구성된다.

이들 망에 인식하고자 하는 음성데이터를 입력하여 각 망의 출력층에 나타나는 평균예측오차를 계산하여 이 값이 최소가 되는 망을 인식망으로 선정하여 인식률을 계산한다.

그림 1의 구조를 가진 다층회귀예측신경망과 4층으로 구성된 다층퍼셉트론(비회귀형)의 인식 실험결과는 표 2와 같다. 각 망 공히 연결세기의 초기치는 0~1.0이고, 하위은닉층의 전달함수는 단극성 시그모이드함수이다. 표2의 인식결과를 비교분석하면 다음과 같다.

① 다층퍼셉트론

망의 구조에 따라 최저 36.07%에서 최대 85%까지 인식률의 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 인식률을 나타내지만 평균적으로 66.05%, 63.38%와 59.51% 정도의 수준이므로 본 연구에서 기대하는 음성인식기로 미흡하다고 판단된다.

② 하위은닉층의 출력을 하위은닉층으로 회귀한 경우;

인식률의 변동폭이 다층퍼셉트론만큼 크지 않고, 평균인식률이 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 신경망은 하위은닉층에서 회귀한 경우보다 평균인식률이 CV음절에서 3.49%, CVC음절에서 1.38%, CV+CVC음절에서 1.6%정도 향상되며, 세 종류의 회귀신경망중 가장 성능이 우수한 신경망이다. 하위은닉층의 뉴런이 10개, 상위은닉층의 뉴런이 10개, 15개일 때 성능이 가장 양호하다.

④ 출력층의 출력을 하위은닉층으로 회귀한 경우;

이 신경망의 인식결과는 하위은닉층 회귀신경망의 결과보다 향상되지만 상위은닉층 회귀신경망에 비해 인식성능이 다소 떨어진다. 2개의 은닉층과 출력층을 거쳐 나감으로써 더욱더 세분된 패턴으로 분류되어 오히려 인식률이 저하된다고 판단된다.

위의 실험결과 상위은닉층의 출력을 하위은닉층으로 회귀할 때 인식성능이 가장 양호하였으므로 이 망을 기준으로 연결세기의 초기치를 변경하고 시그모이드형 전달함수를 변경하면서 인식률의 변화를 알아 보았다. 그 결과는 표3과 같고 비교분석 결과는 다음과 같다.

① 연결세기의 초기치를 0에서 1.0까지 불규칙하게 설정하고, 단극성 시그모이드형 전달함수를 사용하는 경우 :

망의 구조에 따라 인식률이 최저 67.85%에서 최대 90.71%까지 심하게 변하며 일반적으로 상·하위은닉층의 뉴런이 10, 15개, 예측차수 3, 4차에서 CV, CVC, CV+CVC 음절에 대해 각각 84.26~90.71%, 72.14~88.57%, 71.07~83.93%의 인식률 분포를 나타내고 평균적으로 83.14%, 77.43%와 75.17% 정도의 인식 수준을 나타낸다.

② 연결세기의 초기치를 -0.5에서 0.5까지 불규칙하게 설정하고, 단극성 시그모이드형 전달함수를 사용하는 경우 :

인식률의 변동폭은 ①에 비해 최저 인식률은 67.85%에서 71.43%까지 3.58%, 최대 인식률은 90.71%에서 92.14%까지 1.43%가 향상되고 평균인식률 또한 CV음절에서 2.76%, CVC음절에서 1.16%, CV+CVC음절에서 2.83% 상승한다. 특히 상, 하위은닉층의 뉴런이 각각 15개일 때 CV음절인식에서 예

표 2. 다층회귀예측신경망과 다층퍼셉트론의 음성인식률(%)
Table 2. Speech recognition rate of MLRPNNs and MLP(%)

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

표 3. 상위은닉층 회귀예측신경망의 음성인식률(%)

Table 3. Speech recognition rate of the upper hidden layer recurrent prediction neural network(%)

상위 은닉층 뉴런수	하위 은닉층 뉴런수	예측 차수	연결세기 초기치(0~1.0) 단극성 시그모이드함수			연결세기 초기치(-0.5~0.5) 단극성 시그모이드함수			연결세기 초기치(-0.5~0.5) 양극성 시그모이드함수		
			CV	CVC	CV+CVC	CV	CVC	CV+CVC	CV	CVC	CV+CVC
5	5	2	77.86	69.29	68.93	84.29	71.43	71.79	78.57	72.86	71.43
		3	75.00	72.14	67.85	81.43	74.29	73.21	73.57	70.71	67.86
		4	80.00	75.00	71.07	83.57	75.71	73.57	77.76	72.14	67.86
	10	2	82.14	72.14	73.21	83.57	71.43	73.57	83.57	70.00	73.93
		3	85.00	73.57	75.00	83.57	72.86	74.29	79.29	72.86	72.86
		4	83.57	75.71	74.29	84.29	77.14	77.50	80.71	75.71	74.64
	15	2	83.57	75.00	74.64	84.29	72.86	73.93	84.29	75.00	74.29
		3	85.00	73.57	73.57	84.29	70.00	72.50	83.57	74.29	74.64
		4	86.43	77.86	77.86	84.29	77.86	77.50	85.00	81.43	80.71
10	5	2	82.86	72.86	72.86	80.71	72.14	71.79	77.86	71.43	69.29
		3	80.00	76.43	72.86	82.14	76.43	72.50	77.86	72.14	68.57
		4	80.00	76.43	72.86	82.86	74.29	72.89	78.57	72.14	69.64
	10	2	84.26	75.71	76.79	87.14	76.43	79.29	86.43	83.57	81.43
		3	86.43	78.57	77.14	87.86	80.71	78.93	84.29	84.29	82.14
		4	87.14	82.86	80.71	90.71	82.86	83.21	86.43	82.86	80.71
	15	2	85.00	86.43	82.14	89.29	87.86	84.64	87.86	88.57	84.29
		3	90.71	83.57	83.93	91.43	88.57	87.79	86.43	85.00	83.93
		4	85.00	88.57	81.76	88.57	84.29	82.86	85.71	86.43	83.93
15	5	2	84.29	80.00	72.50	82.14	74.29	72.86	78.57	70.00	68.93
		3	79.29	74.29	70.00	77.86	74.29	70.36	78.57	72.86	70.36
		4	82.14	77.86	75.00	85.00	76.43	76.07	80.00	68.57	69.29
	10	2	81.43	73.57	70.71	85.71	84.29	81.07	90.71	88.57	84.64
		3	86.43	80.71	77.14	91.43	83.57	83.93	81.43	82.86	78.21
		4	80.00	72.14	71.07	89.29	83.57	82.86	85.57	84.29	81.07
	15	2	82.14	76.43	75.00	92.14	87.14	86.07	91.43	91.43	89.29
		3	85.00	87.14	82.86	91.43	84.29	85.36	90.00	91.43	87.50
		4	84.26	82.86	77.86	90.00	87.14	85.71	84.29	86.43	82.86
평균인식율			83.14	77.43	75.17	85.90	78.59	78.00	82.90	78.81	76.82

측차수에 관계없이 90%이상의 인식률을 나타내며 상위은닉층의 뉴런이 15개이고 하위은닉층의 뉴런이 10개, 예측차수 4차인 경우, 그리고 동일 상위은닉층에서 하위은닉층 뉴런 15개, 예측차수 2차일 때 ①에 비해 CV, CVC, CV+CVC에서 9.28~11.79%의 인식률 향상이 나타난다. 따라서 연결세기의 초기치가 경우에 따라 상당한 영향을 준다는 것을 알 수 있다.

③ 연결세기의 초기치를 -0.5에서 0.5까지 불규칙하게 설정하고 양극성 시그모이드형 전달함수를 사용하는 경우 :

이 방법은, ①에 비해 ②에서 인식률이 향상되었으므로 ②의 단극성 시그모이드를 양극성 시그모이드로 변환하였을 때 인식률의 변화를 알아보기 위한 실험이다.

인식결과는 CV음절에서 평균인식률이 82.90%, CVC음절에서 78.81%, CV+CVC음절에서 76.82%인데 이것은 ①에 비해 평균인식률이 각각 -0.24%, 1.38%, 1.65%, ②에 비해 평균인식률이 각각 -3%, 0.22%, -1.18% 차이가 있다. 따라서 CVC음절에 대해서만 ①, ②에 비해 어느정도 인식률 향상을 가져오지만 CV, CV+CVC음절에 대해서는 오히려 인식률이 저하된다. 따라서 시그모이드함수는 인식률 향상에 크게 영향을 미치지 않는 다는 것을 알 수 있다.

Ⅳ. 결 론

4층구조의 다층회귀예측신경망의 음성인식성능을 알아보기 위해 4층의 다층퍼셉트론구조에서 하위은닉층, 상위은닉층과 출력층의 출력을 각각 하위은닉층으로 귀환하는 3 종류의 다층회귀예측신경망을 구성하고 각 망에 대한 음성인식성능을 실험을 통해 비교분석해 보았다.

실험결과 회귀형 신경망은 비회귀형에 비해 음성인식성능이 우수하고 구성한 3 종류의 회귀예측신경망중에서 상위은닉층의 출력을 하위은닉층으로 회귀하는 구조에서 상·하위은닉층의 뉴런을 10 혹은 15개, 예측차수를 3 혹은 4차로할 때 비교적 양호한 인식기로 동작한다는 것을 알 수 있었다.

연결세기의 초기치를 0~1.0사이의 불규칙한 값으로 설정할 때 보다 -0.5~0.5사이의 불규칙한 값으로 설정할 때, 그리고 양극성 시그모이드형 함수보다 단극성 시그모이드형 함수를 사용할 때 인식률이 향상되었다.

참고문헌

[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, No. 3, 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] 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 Reseach in Languge, University of California, San Diego, La Jolla, California, Apr. 1988.

[5] 유세관, 나경민, 임재열, 안수길, "회귀신경예측모델을 이용한 음성인식", 전자공학회 하계종합학술대회 논문집, 제18권 1호, pp. 1114-1118, 1995.

[6] 어태경, 배송학, 김주성, 안점영, "다층회귀신경망을 이용한 음성인식", 제15회 음성통신 및 신호처리 워크샵논문집, KSCSP'98, Vol. 15, No. 1, pp. 267-271, 1998.

안 점 영 (Jeom-Young Ahn)

1964년 2월 한국항공대학교 응용전자공학과 졸업

1979년 2월 동아대학교 대학원 전자공학과 졸업(공학석사)

1986년 8월 동아대학교 대학원 전자공학과 졸업(공학박사)

1987년 3월 ~ 현재 동의대학교 전기·전자공학부 교수

* 주관심 분야 : 음성인식, 신경망, 디지털신호처리