2단 회귀신경망의 숫자음 인식에 관한 연구

정회원 안 점 영*, 김 영 재*, 허 강 인**

A Study on the Spoken Digit Recognition Performance of the Two-Stage Recurrent Neural Network

Jeom Young Ahn*, Young Jae Kim*, Kang In Hur** *Regular Members*

요 약

은닉층과 출력층의 신호를 각각 은닉층으로 귀환하는 2단 회귀구조를 갖는 신경망을 구성하고, 이 신경망의 음성인식성능을 평가하기 위해 은닉층의 뉴런수, 입력데이터의 예측차수, 결정상태층의 자기회귀계수를 조정하면서 한국어 숫자음 /공/에서 /구/까지의 음절에 대한 인식실험을 실시하였다.

실험결과, 이 신경망의 인식률은 화자종속의 경우 91%~97.5%, 화자독립의 경우 80.75% ~92%로 나타났고, 이 성능은 화자종속의 경우 Jordan망과 Elman망의 인식수준과 비슷하지만, 화자독립의 경우 다소 우세하다.

ABSTRACT

We compose the two-stage recurrent neural network that returns both signals of a hidden and an output layer to the hidden layer. It is tested on the basis of syllables for Korean spoken digit from /gong/ to /gu/. For these experiments, we adjust the neuron number of the hidden layer, the predictive order of input data and self-recurrent coefficient of the decision state layer.

By the experimental results, the recognition rate of this neural network is between 91% and 97.5% in the speaker-dependent case and between 80.75% and 92% in the speaker-independent case. In the speaker-dependent case, this network shows an equivalent recognition performance to Jordan and Elman network but in the speaker-independent case, it does improved performance.

I. 서 론

인간이 자연스럽게 발음한 음성을 인식하기 위해서는 음성 상에 존재하는 여러 가지 변이 즉, 화자간의 음성학적 차이와 발성시간의 다양성 그리고 조음현상 등의 처리가 잘 이루어져야 한다. 음성인식과 같이 동적 패턴정합 문제를 해결할 수 있는 방법의 하나로 신경망 모델이 있다. 신경망은 사람의 뇌조직을 모방한 것으로서 분산된 기억체계, 병렬신호처리, 반복에 의한 학습과 같은 특징을 가지고 있으며, 이 중에서 주어진 정보들에 내재되어 있는 특성을 학습하여 일반화시킬 수 있다는 것이 신경망의 가장 큰 특징이다.[1]

신경망 중에서 회귀 신경망은 다층 퍼셉트론(Multi-layered Perceptron)에 회귀연결을 부가한 것으로서 문맥 정보를 나타낼 수 있는 내부 상태를 은닉층으로 회귀하므로 동적인 비선형성을 갖는 시스템의 모델링에 적합하고, 회귀연결 구조가 시간적인 변이를 흡수할 수 있으므로 시간축 정렬 알고리듬이 불필요하다.[2][3]

* 동의대학교 전기·전자·정보통신공학부
** 동아대학교 전기·전자·컴퓨터공학부
논문번호: 99243-0614, 접수일자: 1999년 6월 14일

Ⅱ. 2단 회귀신경망의 구성 및 학습

2.1. 망의 구성

본 연구를 위하여 구성한 2단 회귀신경망은 그림 1과 같다. 은닉층의 출력을 복사한 신호를 내부상태층(Internal state layer), 그리고 출력층의 출력을 복사한 신호를 결정상태층(Decision state layer)이라 하며, 결정상태층은 1보다 작은 고정세기의 자기회귀루프를 가지고 있다. 2단 회귀연결과 자기회귀루프를 통해 신경망에 동적인 특성이 부여된다.

2.2 학 습

음성데이터는 단독 숫자음 /공/에서 /구/까지의 음성을 분석한 10차의 LPC melcepstrum을 사용하고, 인접 음성데이터간의 동적특성을 최대로 보존하기 위해 이전시간의 음성데이터로부터 현재시간의 음성데이터를 예측한다. 이를 위한 음성패턴은 그림 2와 같다

그림 2(a)에서 $S^p(1), \cdots, S^p(T)$는 분석된 프레임을 의미하고, $s_1(t), \cdots, s_N(t)$, (t =1, 2,…, T)는 프레임당 음성데이터이다.

그림 2(b)에서 $S^p(t)$를 목표출력(교사신호)으로 할 때, 그 이전시간의 신호를 순차적으로 연결한 신호{ $s^p(t-2)$ $s^p(t-1)$ }는 예측차수 2차에 대한 입력이다.

그림 1의 입력층의 각 뉴런에 그림 2(b)와 같은 입력데이터가 제시되면 이 신호는 은닉층으로 전달되어 은닉층의 뉴런을 거치면서 변환되어 최종적으로 출력층으로 나오게 된다.

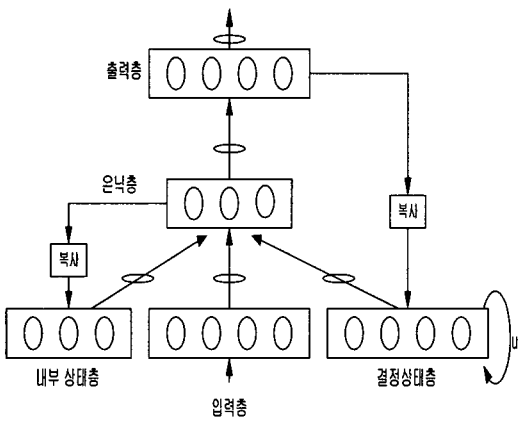
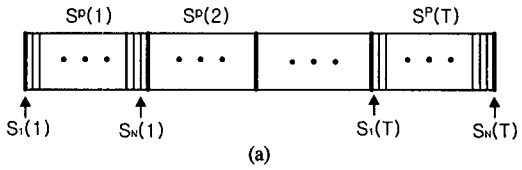


그림 1. 2단 회귀신경망의 구조



(a)

입력신호		교사신호
S^P(t-2)	S^P(t-1)	S^P(t)

(b)

그림 2. (a) 음성패턴
(b) 예측차수 2차의 학습패턴

이 출력(예측신호)과 목표출력(교사신호)을 비교하여 출력층의 자승오차를 계산하고, 이 오차를 감소시키는 방향으로 층 사이의 연결세기를 조절하는 것이 오차역전파 학습이다.[4] 여기서는 오차역전파 학습알고리듬으로, 전향연결만 학습하고, 회귀연결은 학습하지 않는다.

구체적인 학습과정은 다음과 같다. 입력층의 뉴런수를 n개, 은닉층의 뉴런수를 p개, 그리고 출력층의 뉴런수를 q개라 한다. 출력층과 은닉층의 출력이 결정상태층과 내부상태층으로 각각 회귀될 때 은닉층의 j번째 뉴런의 입력성분 $net_j(t)$와 출력성분 $z_j(t)$는 각각 다음과 같다.

$$net_j(t) = \sum_{i=1}^{p+q+n} w_{ji}(t)\, x_i(t), \quad j=1,2,\cdots,p \qquad (1)$$

$$z_j(t) = f(net_j(t))$$

여기서,

$$f(net_j(t)) = \frac{1}{1+\exp(-\lambda\, net_j)} \qquad (2)$$

이다. 위 식(1)에서 $w_{ji}(t)$는 결정상태층, 내부상태층과 입력층에서 은닉층의 j번째 뉴런으로 향하는 연결세기이고, 시간 t에서 예측차수가 τ인 경우 $x_i(t)$는 { $s^p(t-\tau)$ … $s^p(t-1)$ }의 $n\tau$개의 입력데이터와 결정상태층 및 내부상태층 성분을 결합한 벡터이다. 식 (2)는 단극성 시그모이드형 전달함수로서, λ(>0)는 $net_j(t)=0$ 근처에서 연속함수 $f(net_j(t))$의 경사도를 결정하는 뉴런의 이득에 비례한다.

출력층은 선형전달함수를 사용하며, 출력층의 k번째 뉴런의 입력과 출력성분은 각각 다음과 같다.

$$net_k(t) = \sum_{j=1}^{q} w_{kj}(t)\, z_j(t), \qquad k=1,2,\cdots,q \tag{3}$$

$$y_k(t) = \hat{s}_k(t) = net_k(t)$$

$\hat{s}_k(t)$는 예측성분이다. 예측성분이 곧 출력성분이므로 출력층의 k번째 뉴런의 오차성분은 다음과 같다.

$$e_k(t) = s_k(t) - \hat{s}_k(t) = s_k(t) - y_k(t) \tag{4}$$

여기서, $s_k(t)$는 교사신호이다.

따라서, 출력층 전체의 오차는 식 (5)와 같다.

$$E(t) = \frac{1}{2}\sum_{k=1}^{q}[e_k(t)]^2 \tag{5}$$

출력층의 k번째와 은닉층의 j번째 뉴런의 출력오차 δ_{y_k}, δ_{z_j}는 각각 다음과 같다.

$$\delta_{y_k} = s_k(t) - y_k(t) \tag{6}$$

$$\delta_{z_j} = -\frac{\partial E(t)}{\partial net_j(t)} = -\frac{\partial E(t)}{\partial z_j(t)} \cdot \frac{\partial z_j(t)}{\partial net_j(t)}$$

$$= f'(net_j(t)) \sum_{k=1}^{q} \delta_{y_k} w_{kj} \tag{7}$$

연결세기의 변화는 오차값과 뉴런으로 들어오는 입력신호에 비례하고, 또 경사강하시 오차진동을 조절하고, 수렴속도를 증가시키기 위해 상수 α인 모멘텀(momentum)항을 첨가하여 다음과 같은 방법으로 연결세기를 각각 조절한다.

$$w_{kj}(t+1) = w_{kj}(t) + \eta\delta_{y_k} z_j(t) + \alpha\triangle w_{kj}(t-1)$$

$$w_{ji}(t+1) = w_{ji}(t) + \eta\delta_{z_j} x_i(t) + \alpha\triangle w_{ji}(t-1) \tag{8}$$

식 (8)에서 η를 학습계수(learning rate)라 한다. η는 값이 클수록 연결세기가 크게 변화하여 학습속도를 증가시키지만 신경망이 수렴하지 못할 우려가 있다. 반면에 η가 너무 작으면 국부진동은 방지되지만 학습속도가 늦어지는 문제점이 있다.

결정상태층의 i번째 뉴런의 출력신호는 출력층에서 회귀된 신호와 결정상태층의 이전시간 신호를 더한 값으로 식 (9)와 같다.

$$s_d(t) = y_k(t-1) + \mu s_d(t-1)$$

$$= \mu^t s_d(0) + \sum_{n=1}^{t} \mu^{n-1} y_k(t-n) \tag{9}$$

식 (9)에서 μ는 결정상태층의 자기회귀계수이고, $s_d(0)=0$으로 한다.

μ의 값이 1에 가까울수록 결정 상태층의 활성화 레벨이 시간축에서 천천히 변하며, μ=1이면 단순합의 기능을 수행하고, μ=0이면 바로 직전 출력상태의 영향만 나타난다.

Ⅲ. 실험결과 및 고찰

3.1 음성데이터

남성화자 20명이 한국어 단독 숫자음 /공, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구/ 10개를 각각 4회씩 발성한 총 800개의 음성을 상한주파수 7kHz의 LPF를 거친 후 16kHz로 샘플링하여 16bit로 양자화한다. 이 신호를 고역강조한 후 16ms의 Hamming창으로 3.75ms씩 이동하면서 LPC cepstrum계수를 추출하고, 이를 다시 10차의 melcepstrum 계수로 변환하여 음성데이터로 사용한다.

3.2 실험

남성화자 20명이 각각 4회씩 발성한 음성데이터 중에서 10명의 2회분(200개) 음성데이터는 학습용으로, 그리고 동일화자의 2회분(200개)은 화자종속 실험용으로 사용하였다. 나머지 10명의 음성데이터 400개는 화자독립실험에 이용하였다. 연결세기는 -0.5에서 0.5사이의 랜덤수로 초기화[5]하고, 모멘텀 항 계수 α는 0.9, 학습계수는 0.0001로 고정시켰다. 또한 경사도 λ 는 1로 하였다. 학습은 3000회까지 실행하고, 학습이 종료되면 총 10개의 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 계산하였다. 구성된 2단 회귀신경망의 인식성능을 알아보기 위해 은닉층의 뉴런수를 10, 15, 20, 25개로 변경하고, 입력신호의 예측차수를 2, 3, 4차로 조정하며, 결정상태층의 자기회귀계수 μ를 0, 0.4, 0.5, 0.6으로 변화시켰다. 그리고 이 망의 출력층의 뉴런수는 10개이다.

또한, 제안된 회귀신경망의 인식성능 수준을 객관적으로 평가하기 위해 일반적으로 많이 알려진 Jordan망[6]과 Elman망[7][8]의 인식성능과 비교하였다.

3.3 고찰

2단 회귀신경망의 음성인식 결과는 표 1과 같다.

표 1을 보면 2단 회귀신경망은 자기회귀계수, 예측차수, 은닉층의 뉴런수에 따라 인식률이 다양하게 변화하며, 화자종속인 경우 예측차수 2차, 은닉층의 뉴런수 20개, 자기회귀계수 μ가 0일 때 97.5%, 화자독립인 경우, 예측차수 3차, 은닉층의 뉴런수 10개, 자기회귀계수 μ가 0일 때 92%로 가장 높은 인식률을 보였으며, 자기회귀계수 μ값이 증가할수록 인식률이 대체로 열화됨을 볼 수 있었다.

표 1. 2단 회귀신경망의 자기회귀계수, 예측차수, 은닉층의 뉴런수 변화에 따른 인식률(%)

예측 차수	*은닉 층수	u=0		u=0.4		u=0.5		u=0.6	
		종속	독립	종속	독립	종속	독립	종속	독립
2	10	92.5	85.75	93.50	83.25	91.50	86.50	91.50	81.25
	15	93.00	89.50	94.00	87.75	94.00	88.00	93.00	84.75
	20	97.50	89.50	93.50	87.00	94.00	87.50	93.50	85.00
	25	93.00	88.75	94.00	90.50	94.50	88.00	94.00	87.50
3	10	96.50	92.00	95.00	88.25	92.00	91.00	95.00	86.25
	15	95.00	88.75	94.00	89.75	93.50	85.5	94.00	87.75
	20	95.50	89.75	96.50	91.50	95.50	86.25	94.50	89.50
	25	93.50	88.75	95.50	87.50	94.00	90.25	96.00	85.50
4	10	93.50	85.75	90.00	81.75	91.50	85.00	90.50	80.75
	15	95.50	87.00	93.00	86.75	91.00	84.25	92.00	85.75
	20	94.00	85.50	93.00	87.75	92.50	89.50	93.00	85.50
	25	92.50	86.25	95.50	87.50	94.50	88.00	94.50	84.50

* 은닉층수는 은닉층의 뉴런수임

인식결과를 구체적으로 알아보기 위하여 비교적 인식률이 좋은 은닉층의 뉴런수 10개, 결정상태층의 자기회귀계수 0, 입력신호의 예측차수 3차의 구조에서 각 음성별 오인식률을 조사하였다.

표 2. 화자종속인 경우 2단 회귀신경망의 오인식 발생회수

A \ B	공	일	이	삼	사	오	육	칠	팔	구
공	19									1
일		19						1		
이			20							
삼				20						
사					20					
오	1					16				3
육							20			
칠								20		
팔									20	
구	1									19

A:인식대상음성, B:인식음성

표 2는 학습에 참여한 10명에 대한 화자종속 인식결과이다. 이 경우 음절당 인식대상음성은 각각 20개이며, 표 1에 의하면 인식률은 96.5%이다. /공/과 /구/에서 오인식이 상호 1회씩 발생하였으며, 특히 /오/가 /구/로 3회씩이나 오인식되어 전체인식률에 크게 영향을 주고 있음을 알 수 있다.

표 3은 학습에 참여하지 않은 20명에 대한 화자독립 인식결과이다. 이 경우 음절당 인식대상음성은 각각 40개이며, 인식률은 92%이다. 화자종속인 경우에 비해 여러 음성에서 오인식이 발생하였다. 특히, /공/이 /오/와 /구/로, /일/이 /이/와 /칠/로, /오/가 /공/과 /구/로, 그리고 /구/가 /공/으로 오인식이 많이 된다는 것을 알 수 있었다.

표 3. 화자독립인 경우 2단 회귀신경망의 오인식 발생회수

A \ B	공	일	이	삼	사	오	육	칠	팔	구
공	32					2	1		1	4
일		31	2					7		
이		2	37				1			
삼				38	2					
사					40					
오	4					34				2
육							40			
칠								40		
팔	1								39	
구	2					1				37

A:인식대상음성, B:인식음성

표 2와 3에 나타난 /공/,/오/,/구/ 그리고 /일/, /이/, /칠/의 오인식을 줄일 수 있다면 전체적으로 인식률이 크게 향상될 것이다.

표 4는 2단 회귀신경망과 동일한 조건으로 실험한 Jordan망과 Elman망의 인식결과이다. 2단회귀신경망의 자기회귀계수 μ=0일 때의 인식률과 비교하기 위해 Jordan망 역시 μ=0으로 하였다.

실험결과, 각 망의 크기에 따라 다양한 인식률을 보였으며, Jordan망의 경우 예측차수 4차, 은닉층의 뉴런수 25개일 때 화자종속에서 97.5%, 예측차수 3차, 은닉층의 뉴런수 20, 25개일 때 화자독립에서 88.75%의 최대인식률을 보였다. 또한, Elman망의 경우 예측차수 3차, 은닉층의 뉴런수 20개일 때 화자종속에서 96.5%, 화자독립에서 89.5%의 가장 좋은 인식률을 나타내었다.

표 4. Jordan망과 Elman망의 인식률(%)

예측 차수	*은닉 층수	Jordan(u=0)		Elman	
		종속	독립	종속	독립
2	10	94.50	85.75	93.50	83.25
	15	93.00	87.50	94.00	87.75
	20	95.50	88.50	93.50	87.00
	25	96.00	88.25	94.00	88.50
3	10	94.50	87.00	95.00	87.25
	15	94.00	88.25	94.00	88.00
	20	95.50	88.75	96.50	89.50
	25	95.00	88.75	95.50	87.50
4	10	92.00	85.75	90.00	81.75
	15	93.00	87.00	93.00	86.75
	20	93.00	85.50	93.00	87.50
	25	97.50	85.25	95.50	87.50

* 은닉층수는 은닉층의 뉴런수임

제안된 2단 회귀신경망의 인식성능을 Jordan망과 Elman망의 인식성능과 객관적으로 비교해 볼 때 화자종속인 경우에는 비슷한 수준이지만, 화자독립인 경우에는 2단 회귀신경망의 성능이 다소 우세하다는 것을 알 수 있다.

Ⅳ. 결 론

입력층, 은닉층과 출력층으로 구성된 다층 퍼셉트론에서 은닉층과 출력층의 출력값을 각각 은닉층으로 회귀하는 2단 회귀신경망을 구성하고, 이 신경망의 음성인식성능을 평가하기 위해 은닉층의 뉴런수, 입력 데이터의 예측차수, 결정상태층의 자기회귀계수를 각각 변화시키면서 /공/에서 /구/까지의 단독숫자음에 대한 인식실험을 수행하였다.

실험결과, 은닉층의 뉴런수, 예측차수, 자기회귀계수에 따라 화자종속 인식률은 91%~97.5%, 화자독립 인식률은 80.75%~92%의 수준이었다. 이 인식률은 화자종속의 경우 Jordan망과 Elman망의 인식률과 대체로 비슷하지만, 화자독립인 경우에는 두 신경망보다 상승된 결과이다.

참 고 문 헌

[1] 유제관, 나경민, 임재열, 안수길, "회귀 예측모델을 이용한 음성인식", 전자공학 회 하계종합학술대회 논문집, 제18권 1호, 1995

[2] 김기석, 황희웅, "순환신경망 모델을 이용한 한국어 음소의 음성인식에 대한 연구", 전기학회 논문지, 제40권 8호, pp.782-791, 1991

[3] J. Ludik, W. Prins, K. Meert, T. Catfolis, "A Comparative Study of Fully Partially Recurrent Networks", Proceeding of ICNN 97, Vol.1, pp292~297, 1997

[4] Jacek M. Zurada, "Introduction to Artificial Neural Systems", West Publishing Company, pp.175~213, 1992.

[5] 어태경, 안점영, "연결 세기 초기치와 시그모이드 함수 변화에 따른 상위은닉층 회귀신경망의 음성인식성능 비교", 동의논집 제30집 자연과학편, pp. 627 ~632, 1999

[6] Micheal L. Jordan., "Attractor dynamics and parallelism in a connectionist sequential machine", Proc. for the 1986 Cognitive Science Conference, pp. 531-546, 1986

[7] Ernst Haslsteiner, "What Elman Networks cannot do", IJCNN 98, Vol.2, pp.1245-1249, 1998

[8] D.T. Pham, X. Liu, "Training of Elman networks and dynamic system modelling", International Journal of Systems Science, Vol.27, number 2, pp.221~226, 1996

안 점 영(Jeom-young Ahn) 정회원

현재 : 동의대학교 전기 · 전자 · 정보통신공학부 교수

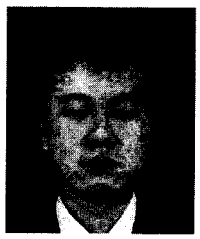
제24권 제12B호 참조

허 강 인(Kang-in Hur) 정회원

현재 : 동아대학교 전기 · 전자 · 컴퓨터공학부 교수

제24권 제12B호 참조

김 영 재(Young-jae Kim) 정회원



1998년 2월 : 동의대학교 공과대학 전자공학과(공학사)

2000년 2월 : 동의대학교 대학원 전자공학과(공학석사)

<주관심 분야> 음성신호처리, 디지털 신호처리