醫工學會誌 : 第11卷, 第1號, 1990
J. of KOSOMBE : Vol. 11, No. 1

研 究 論 文
90-1-20

청각모델과 회귀회로망을 이용한 음성인식에 관한 연구

김 동 준 · 이 재 혁 · 윤 태 성* · 박 상 희

A Study on Speech Recognition Using Auditory Model and Recurrent Network

Dong-Jun Kim, Jae-Hyuk Lee, Tae-Sung Yoon*, and Sang-Hui Park

-Abstract-

In this study, a peripheral auditory model is used as a frequency feature extractor and a recurrent network which has recurrent links on input nodes is constructed in order to show the reliability of the recurrent network as a recognizer by executing recognition tests for 4 Korean place names and syllables.

In the case of using the general learning rule, it is found that the weights are diverged for a long sequence because of the characteristics of the node function in the hidden and output layers. So, a refined weight compensation method is proposed and, using this method, it is possible to improve the system operation and to use long data. The recognition results are considerably good, even if time worping and endpoint detection are omitted and learning patterns and test patterns are made of average length of data. The recurrent network used in this study reflects well time information of temporal speech signal.

1. 서 론

인간의 기능에 대한 끊임없는 탐구는 많은 발전을 이루어 왔다. 그 중에서도 인간의 언어에 대한 기계의 인식을 목적으로 하는 음성인식에 관한 연구는 활발히 진행되어 왔으나, 뇌에서의 인식기능에 대한 연구는 최근에 비로소 활발해졌다. 뇌에서의 인식은 공학적인 어떤 방법보다 탁월한 능력을 나타낸다. 신경회로망을 음성인식에 적용한 연구도 최근에 상당히 발표되었다. 대부분의 연구가 음성의 시간적인 정보를 충분히 반영하지 못하는 정적인 패턴에 대해 적용 가능한 정적인 신경회로망을 이용하였다.

본 연구에서는 음성의 주파수 특징을 추출하기 위해 말초 청각모델을 이용하고, 인식단에서 음성의 시간정보를 그대로 제공할 수 있는 회귀회로망을 이

〈접수 : 1990년 5월 7일〉
연세대학교 전기공학과
*창원대학교 전기공학과
Dept. of Electrical Engineering, Yonsei University
*Dept. of Electrical Engineering, Changwon University

용하여 한국어 단어 및 음절에 대한 인식실험을 하였다. 이용된 회귀회로망은 입력단에만 회귀링크가 있어서 입력정보의 내력을 나타낼 수 있게 한 구조이다. 이 모델을 이용하여 학습실험에 있어서의 동작특성과 문제점을 지적하여 이에 대해 해결할 수 있는 방법을 모색하였으며, 또한 문맥 정보가 충실히 반영되는 상태에서 대신 시간워핑을 포함한 과거의 정교한 전처리과정을 간단하게 줄여서 모델동작에 대한 신뢰성을 확인하고자 한다.

2. 말초 청각 계통 모델과 회귀회로망

2.1 말초 청각 계통 모델

본 연구에서는 음성신호의 분석 및 인식처리에 활용할 목적으로 하는 말초 청각 계통의 각 부분에 대한 모델을 실험 모델로 사용하였다. 외이와 중이의 특성을 고려한 Monro의 외이의 2차 모델과 중이의 1차 모델을 이용하였다.[1] 대역통과 필터와 같은 전달특성을 갖는 내이의 경우, 중심주파수가 300[Hz] 이하에서는 Yegnanarayanan의 2차 공진 모델을, 300[Hz] 위에서는 Flanagan의 모델을 사용하였다.[2,3]

헤어셀/청각신경 모델은 Schroeder와 Hall의 모델을 이용하였다.[4]

2.2 회귀회로망

음성에 대한 정보는 시간에 걸쳐서 복잡한 방법으로 분포되어 있다. 시간에 걸쳐 분포된 시이퀀스의 특징의 포착을 위해서는 시이퀀스의 상태정보를 회로망에 제공해야 하며, 이때 회귀성(recurrent), 혹은, 피이드백(feedback) 링크가 필요하다. 이는 시간적인 패턴의 처리에 있어서 기대하는 출력을 얻기 위해서는 현재 입력의 앞과 뒤의 입력도 큰 영향을 미치기 때문이다. PDP그룹(1977), Jordan(1986), 및 Elman(1988) 등에 의하여 여러가지 형태의 회귀회로망이 제시되었다.[5] 이러한 시스템은 매우 고무적인 결과를 냈지만, 고유의 결점을 가지고 있다. 첫째, 각각의 시간 단계를 표현하기 위한 구조의 복제로 인하여 불필요한 계산이 수행되는 것이다. 둘째, 어떤 고정된 길이의 시간적 윈도우를 갖을 경우에는 윈도우의 크기보다 긴 시간적 영향을 검출하는 것은 불가능하다. 셋째, 공간적 표현에서 배열이 약간 다르게 된 신호가 제대로 배열된 학습 신호에 대하여 오인식을 할 가능성이 크다.

본 연구에서는 각 시간 단계에 구조의 복제를 하지 않고 입력 패턴의 시간적 정보를 포착하는 시스템을 사용한다. 고전적인 back propagation algorithm에서 입력노드는 단순히 입력된 신호를 다음층에 분배하는데에만 사용되었는데, 이 시스템에서는 추가적인 계산을 수행하여야 한다.[6]

X_i를 시간 t_i에서 입력 노드에 의해 계산된 값이라 하고, I_i는 같은 시각에 이 노드에의 입력 신호라 하자, 그러면 입력 노드는 다음 식에 의해 연속되는 값을 계산한다.

$$X_{i+1} = aI_{i+1} + dX_i \qquad (2.1)$$

a : 입력의 크기
d : 감쇄율

뒤따르는 입력이 더이상 없을 때에는 입력 노드의 활성화(activation)값은 지수함수적으로 감소한다. 원하는 특정 τ시간 이후에 1/e이 되도록 이 감쇄율 값을 선택한다. 즉 $d = e^{-r\tau}$이고, r은 입력신호 제공률이다. a값은 $aI_{max}/(1-d)$에 의해 계산되는 X의 값이 특정 최대값을 갖도록 선택된다. 여기서, 은닉층과 출력층은 고전적 back propagation net와 같은 구조이다.

2.3 회귀회로망의 학습 알고리즘

고전적인 back propagation algorithm은 다음의 5단계로 이루어진다.

제 1 단계 : 모든 링크의 가중치(weight)와 노드의 옵셋(offset)을 임의의 작은 값으로 초기화한다.

제 2 단계 : 하나의 입력벡터 $X=[x_0, x_1, \cdots, x_{N-1}]$와 이 때의 기대 출력 벡터 $D=[d_0, d_1, \cdots d_{M-1}]$를 제공한다.

제 3 단계 : 입력의 가중치 합을 식(2.2)에 의하여 구하고 식(2.3)의 sigmoid 비선형 함수에 적용하여 출력을 계산한다.

$$net_j = \sum_{i=0}^{n-1} W_{ij} O_i \tag{2.2}$$

$$O_j = f(net_j) = \frac{1}{1+e^{-(net_j-\theta_j)}} \tag{2.3}$$

여기서 W_{ij}는 노드 i와 j의 가중치이고, O_i는 노드 i의 출력 값이고, θ_j는 노드 j의 고유 옵셋이며, 출력층의 실제출력을 $Y=[y_0, y_1, \cdots y_{M-1}]$이라 한다. 출력층에서는 $O_i=y_i$이고, 입력층에서는 $O_i=X_i$이다.

제 4 단계 : 식(2.4)에 의해 가중치 W_{ij}를 조정한다.

$$W_{ij}(t+1) = W_{ij}(t) + \eta \delta_j O_i \tag{2.4}$$

여기서 t는 W_{ij}의 조정이 반복된 횟수를 나타내고, O_i는 노드 i의 출력인데, 입력층에서의 경우는 망입력이다. η는 이득항(gain term) 혹은 학습률(learning rate, $0<\eta<1$)이고, δ_j는 노드 j의 오차항(error term)으로서 노드 j가 출력노드일 때

$$\delta_j = y_j(1-y_j)(d_j-y_j) \tag{2.5}$$

이고, 여기서, d_j는 노드 j의 기대 출력이고 y_j는 실제 출력이다. 만일 노드 j가 내부의 은닉층의 노드라면

$$\delta_j = O_j(1-O_j) \sum_k \delta_k W_{jk} \tag{2.6}$$

이고, 여기서, k는 상위층의 모든 노드에 해당하는 첨자이다. 여기에 모멘텀 항(momentum term)이 추가되면 수렴 속도가 더 빨라지기도 하고, 가중치 변화가 완만해진다.

$$W_{ij}(t+1) = W_{ij}(t) + \eta \delta_j O_i + \alpha(W_{ij}(t) - W_{ij}(t-1)) \tag{2.7}$$

여기서, α는 모멘텀 항이며, $0<\alpha<1$이다. α는 현재의 가중치 변화에 대한 정보를 다음의 가중치 조정에 반영해 주므로, 적당히 클 때 수렴 속도가 빠르다.

제 5 단계 : 모든 입출력 벡터쌍에 대한 실제 출력 벡터와 기대 출력 벡터간의 오차 E가 충분히 작지 않으면 제2단계로 되돌아가 제5단계까지의 과정을 반복하다가 오차가 충분히 작으면 반복을 그친다. 각 패턴에 대한 오차는

$$E_p = \frac{1}{2} \sum_k (d_{pk} - y_{pk})^2 \tag{2.8}$$

d_{pk} : p패턴의 출력단 k번째 노드 기대 출력
y_{pk} : p패턴의 출력단 k번째 노드 실제 출력

이고, 평균 시스템 오차는

$$E = \frac{1}{2p} \sum_p \sum_k (d_{pk} - y_{pk})^2 \tag{2.9}$$

에 의해 계산된다.

본 논문에서는 다층 인식자(multi-layer perceptron)의 학습 규칙인 위의 알고리즘을 회귀회로망의 동작 특성을 고려하여 약간의 수정을 하였다.

먼저, 제 2 단계에서는 하나의 입력 벡터대신 시간에 따라 $x_0, x_1 \cdots, x_{N-1}$ 각각에 시이퀀스가 식(2.1)에 의해 입력되고, 이 때의 기대 출력은 앞에서와 마찬가지로 벡터 $D=[d_0, d_1 \cdots, d_{M-1}]$를 제공한다. 따라서, 입력 패턴은 하나의 행렬을 형성한다. 제 3 단계에서의 입력벡터에 따른 출력이 같은 방법으로 계산된다. 제 4 단계에서는 각 시간 단계마다 가중치 보정을 하지 않고, 전체 시이퀀스의 가중치 변화량을 계산하고 모두 합하여 가중치 조정을 하게 된다. 즉,

$$\eta \delta_j O_j = \angle W_{ij} = \sum_{s=0}^{Q-1} dW_{ij} \tag{2.10}$$

이고, Q는 시이퀀스의 길이를 나타내고, dW_{ij}는 각 시간 단계에서의 가중치 변화량을 나타낸다. 제 5 단계에서 각 시간 단계에서의 오차는

$$E_{pt} = \frac{1}{2} \sum_t (d_{pkt} - y_{pkt})^2 \tag{2.11}$$

d_{pkt} : 시간 단계 t에서 p패턴의 출력단 k번째 노드 기대 출력
y_{pkt} : 시간 단계 t에서 p패턴의 출력단 k번째 노드 실제 출력

이고, 한 패턴에 대한 오차는

$$E_p = \frac{1}{2} \sum_k \sum_t (d_{pkt} - y_{pkt})^2 \tag{2.12}$$

이며, 평균 시스템오차는

$$E=\frac{1}{2p}\sum_{p}\sum_{k}\sum_{t}(d_{pkt}-y_{pkt})^2 \qquad (2.13)$$

에 의해 계산되었다.

3. 실험 및 결과 고찰

3.1 전체 시스템과 실험장치

본 연구에서는 귀에 들어온 음성신호가 청각계통을 거쳐 뇌에서 인식되기까지의 모든 과정을 구현하기 위하여 그림 3.1과 같이 음성 인식 시스템을 구현하였다.

입력단의 노드 수는 헤어셀 출력단의 출력 수와 같이 18개의 노드를 이용하였고, 은닉층은 20개 노드를 이용하였고, 출력단은 인식해야 할 음성패턴의 수에 맞추었다. 본 연구에서는 음성신호의 주파수 대역을 4.7[KHz]로 제한하였고 내이채널의 중심 주파수의 범위를 50~4,000[Hz]로 제한하였다.

음성 데이타로는 한국어 지명을 4인의 남성 화자가 3회 반복 발음을 하여, 4(지명)×4(화자)×3(발음)=48개의 데이타를 이용하였고, CV형태 음절은 2인의 남성 화자가 3회 반복 발음한 총 4(음절)×2(화자)×3(발음)=24개의 데이타를 이용하였다.

3.2 지명의 인식

그림 3.2는 지명 /부산/을 청각모델에 적용하여 18개 채널 출력을 나타낸 것이다.

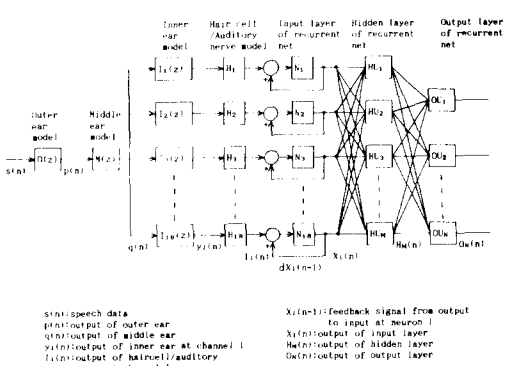

그림 3.1 음성 인식 시스템

Fig. 3.1 Speech recognition system

그림 3.2와 같은 진폭, 채널(주파수), 시간의 3차원 정보를 회귀회로망에 적용하여 학습을 시키기 위하여 각 채널에서 50[msec]로 평균을 취하였다. 이는 회귀회로망의 정보 반영 능력을 인식실험 결과를 통하여 진단해 보고자 한 것이다.

학습률 η는 0.5로 고정하였고, 모멘텀 항 α는 0.9를 선택하였다. η는 너무 작으면 가중치의 수렴 속도가 늦어지고, 큰 값을 선택했을 때는, 수렴이 된다면 그 속도가 빠르지만 발산하는 경우가 많았다.

전체 학습의 수에 대한 시스템 오차의 추이를 그림 3.3에 나타내었다. 그림 3.4와 3.5는 화자종속실험과 화자독립실험에서 4개 출력노드 각각에서 전체 시이퀀스가 회로망을 통과하였을 때 그 출력을 나타낸 것이다.

인식실험의 결과는 표 3.1에 나타나 있다.

화자독립실험에서는 화자 KDH의 발음 /마산/이 /부산/으로 인식되었는데, 이는 두 패턴이 비교적 상당히 비슷한 형태를 띄며, 특히 뒤의 음절은 완전

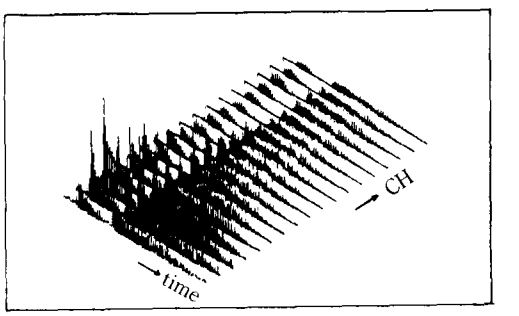

그림 3.2 지명 /부산/의 18채널 청각모델 출력

Fig. 3.2 Outputs of 18 channel auditory model for place name/Pusan/

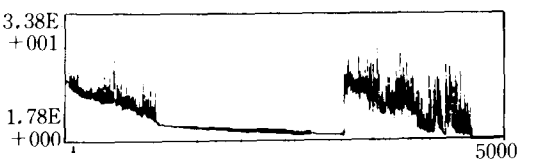

그림 3.3 일반적인 학습방법에 따른 오차곡선

Fig. 3.3 Error curve of general learning method

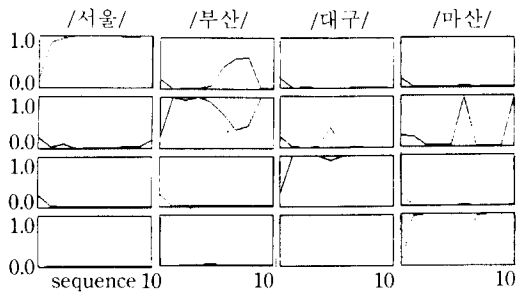

그림 3.4 4개의 지명에 대한 화자종속실험의 결과
Fig. 3.4 Speaker dependent tests on 4 place names

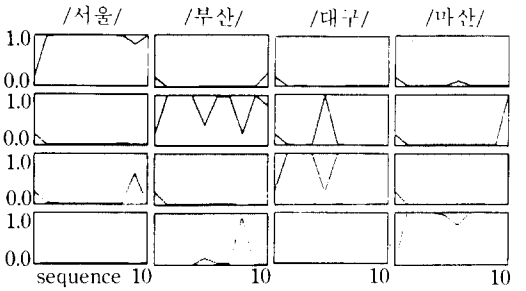

그림 3.5 4개의 지명에 대한 화자독립실험의 결과
Fig. 3.5 Speaker independent tests on 4 place names

히 똑같기 때문에 발생한 현상이다.

3.3 가중치 보정식의 수정

다음에는 평균화의 폭을 줄여서 시이퀀스의 길이를 10배로 늘렸다. 지명은 /서울/과 /마산/ 두 개만을 인식대상으로 하였다.

회귀회로망의 가중치 보정에 대한 일반적인 방법 인식(2.10)을 대입하였을 때 local minima에 빠져버리는 현상이 나타났다. 이러한 현상을 없애고, 시이퀀스 길이에 무관하게 동작할 수 있도록 나름대로의 가중치 보정식을 제안한다.

$$\eta d_j O_i = \angle W_{ij} = \frac{1}{Q} \sum_{s=0}^{Q-1} dW_{ij} \qquad (3.1)$$

Q : 시이퀀스 길이, dW_{ij} : 각 시간단계에서의 가중치 변화량

표 3.1 한국어 지명의 인식실험 결과
Table 3.1 Recognition results of Korean place names

place name	recognition result (number)				error (content)
	speaker dependent	speaker independent			
	KDJ	ASP	KDH	LSJ	
/서울/	3	3	3	3	0
/부산/	3	3	3	3	0
/대구/	3	3	3	3	0
/마산/	3	3	2	3	1(/부산/)
recognition rate(%)	100	100	91.7	100	
		97.2			

이와 같이 하여 계산된 시스템오차를 그림 3.6에 나타내었다.

화자종속실험과 화자독립실험에서 모두 100%의 인식 결과를 얻었다.

3.4 음절 인식실험

앞에서 제시한 가중치 보정식 식(3.1)을 이용하여, CV형태의 음절에 대한 실험을 수행하였다. 자음 /ㄱ/, /ㄴ/과 모음 /ㅏ/, /ㅣ/를 조합한 4개의 음절에 대하여 화자 2명이 3회 발음한 것을 20[KHz]로 샘플링하였다. 데이타는 끝까지 취하지 않고 자음에서부터 2500개(125[msec])를 취하고 12.5[msec]로 평균을 취하였다. 그림 3.7은 전체 학습의 수에 대한 시스템 오차의 추이를 나타낸다. 그림 3.7에서는 오차곡선이 완만한 지수함수모양으로 감소한다.

그림 3.3에서 진동이 심하게 나타났는데, 더구나 시이퀀스가 길면 local minima에 빠지는 결정적인 단점을 안고 있다. 수정한 식(3.1)은 그림 3.7과 같이 비교적 안정된 가중치 보정의 특성을 나타낸다.

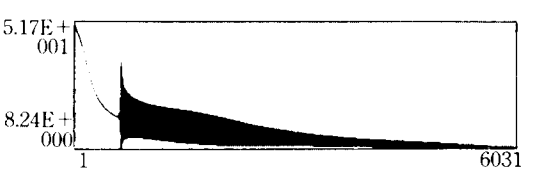

그림 3.6 개선된 보정방식에 따른 오차곡선(Q=100)
Fig. 3.6 Error curve of refined method(Q=100)

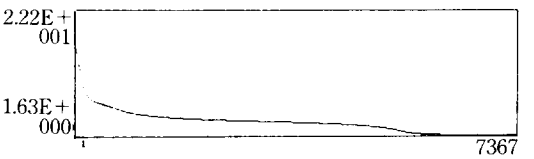

그림 3.7 개선된 보정방식에 따른 오차곡선(Q=10)
Fig. 3.7 Error curve of refined method(Q=10)

표 3.2 한국어 음절의 인식실험 결과
Table 3.2 Recognitin results of Korean syllables

Syllable	recognition result (number)		error (content)
	speaker dependent	speaker independent	
	LSJ	ASP	
/가/	3	3	0
/나/	3	3	0
/기/	3	2	1(/나/)
/니/	3	1	2(/나/)
recog-nition rate(%)	100	75	

인식된 결과는 표 3.2에 나타내었다.

5. 결 론

본 논문에서는 음성신호의 주파수 특징 추출단으로는 말초청각계통모델을 이용하고, 입력단에 회귀링크를 가진 회귀회로망을 구성하여 음성 인식기로서 회귀회로망의 신뢰성을 확인하고자 하였다.

얻어진 결과는 다음과 같다.

1. 본 연구에서 사용한 회귀 회로망을 시간에 따라 입력되는 음성신호의 시간정보를 충분히 반영한다.

2. 회귀회로망에 대한 개선된 가중치 보정방식을 제시하여 실시해 본 결과, 학습에서의 시스템 동작 특성을 개선하고 긴 데이타에도 적용할 수 있게 되었다.

3. 시간 워핑과 끝점 추출을 하지 않고 발음의 길이에 관계없이 어느 일정 길이로 잘라서 학습패턴 및 테스트패턴을 만들어도 상당한 인식 결과를 얻었다. 시간 정보가 잘 반영되면 다른 정보의 큰 손실에도 불구하고 인식기능을 잘 발휘한다고 볼 수 있다.

참 고 문 헌

1) D. M. Monro, "Computer modeling of the peripheral mechanical responce of the auditory system," in Auditory investigation : the scientific and technological basis, edited by H. A. Beaglet, Clarendon Press, pp. 431-450, 1979.
2) G. Yegnanarayanan, "A new model of hearing and its performance in pitch perception," Ph. D. thesis, Delaware Univ., 1985.
3) J. L. Flanagan ; Speech analysis, synthesis and perception, Springer Verlag, pp. 109-112, 1972.
4) M. R. Schroeder and J. L. Hall, "Model for machanical neural transduction in the auditory receptor," JASA Vol. 65(4), pp. 1055-1060, 1974.
5) F. S. Tsung and G. W. Cottrell, "A sequential adder using recurrent networks," IEEE IJCNN, Vol. 1, pp. 133-139, 1989.
6) W. S. Storneta, "A dynamical approach to temporal pattern processing," AIP, 1988.
7) 윤태성, "청각모델을 이용한 한국어 단음의 인식에 관한 연구." 박사학위논문, 연세대학교 대학원, 1988.
8) R.P. Lippman, "An introduction to computing with neural nets," IEEE ASSP Magazine, Vol. 3, No. 4, pp. 4-22, 1987.
9) D. E. Rumelhart, G. E. Hinton and R. J. William ; Learning internal representations by error propagation, in D. E. Rumelhart & J. L. McClelland(Eds.), Parallel Distributed Processing : Explorations in the Microstructure of Cognition, Vol. 1, pp. 318-364, 1986.
10) R. L. Watrous, "Learning phonetic feature using connectionist networks : An experiment in speech recognition," IEEE conf. on Neural Network, Vol. 3, pp. 381-389, 1987.