시공간 패턴인식 신경망에 의한 단어 인식에 관한 연구

박 경 철° 김 헌 기 이 종 호
인하대학교 전기공학과

A Study on Recognition of Spoken Numbers Using Spatio-Temporal Pattern Recognizer

Kyoung Cheol Park Hun Kee Kim Chong Ho Lee
Dept. of Electrical Eng. Inha University

Abstruct
This paper presents spoken numbers recognition method using a spatio-temporal network. This network is efficient in processing the spectrum sequences of speech patterns as spatio-temporal patterns. The number of windows and channels is experimentally determined. The recognition rate has been improved by experiments done on various parameters. The test data is collected from 10 numbers spoken by 2 male and 2 female speakers. A recognition rate of 80% was obtained on a test set of 50 words.

1. 서론

신경회로망을 이용하여 음성인식에 적용한 사례들이 매우 늘어 나고 있다.[1] 물론 음성인식은 매우 어려운 분야이고 이를 위해 해결 해야할 많은 문제점들이 있는 것은 사실 이다. 그러나 인간과 컴퓨터가 자연스럽게 의사 소통을 할 수 있는 음성 인식 시스템을 구성 한다면 그 응용 분야는 거의 무한하다. 과거에도 음성인식을 위하여 많은 시도들이 있었다. DTW(Dynamic Time Warping)나 HMM(Hidden Markov Model) 등의 방법과 전통적인 신경회로망 모델인 역전파 신경회로망,시간의 지연에 따라 신경망을 적용한 TDNN (Time-Delay Neural Network)과 재순환 신경회로망 등의 방법이 있다.[2] 하지만 음성인식을 위한 음성의 정적 또는 동적인 모델링 조차도 충분히 이루어 지지 않았다. 하지만 요즈음 인간의 청각계의 생물학적인 구조가 밝혀지면서 보다 정확한 음성인식 모델을 설계할 수 있게 되었다.[3] 인간은 음성을 인식 하는데 전혀 어려움을 느끼지 않고 자연스럽지만 컴퓨터는 음성을 인식하는데 매우 어려움을 느낀다. 이것은 인간과 컴퓨터가 음성을 인식 하는 인지구조가 전혀 다르기 때문이다.[4] 인간은 음성을 인식하는데 있어서 단순한 음성의 크기변화 보다는 주파수의 변화에 민감 하다.이러한 인간의 생물학적 메커니즘을 모방하는것이 가장 타당하다고 볼수 있다. 이것은 음성의 시간 변동과 주파수 변동을 가장 탁월 하게 처리 하는 것이 인간이기 때문이다.[5] 본 연구에서는 이러한 인지구조를 고려 하여 동적 신경회로망인 시공간 패턴 인식기로 음성인식에 적용 하여 보았다. 남녀 4인이 5회 발음한 10개의 숫자 음을 가지고 음성인식 실험을 하여 여러 파라미타의 최적 값을 구하고 인식률을 검토 하여 보았다.

2. 시공간 패턴인식기

2.1 개요

음성은 시간에 따라 특징이 변화 하며 같은 화자에 의한 발음이라도 발음 속도에 의해 차이가 나며 문맥의 전후 관계에 의해서도 다르게 발음 된다. 그래서 음성과 같이 시간에 따라 순차적으로 변화하는 패턴을 효과적으로 인식하기 위해서는 시퀀스 패턴의 변화를 반영할 수 있는 특징을 가져야 한다. 그러나 역전파 신경회로망 모델이나 홉필드네트워크와 같은 대부분의 모델들은 정적인 패턴을 다루므로 음성인식에 적절하다고 볼 수 없다. 따라서 패턴의 시퀀스를 기억하고 재현할 수 있는 모델이 제안되었는데 이것이 시공간 패턴 인식기이다.[6][7]

2.2 동작 원리와 구조

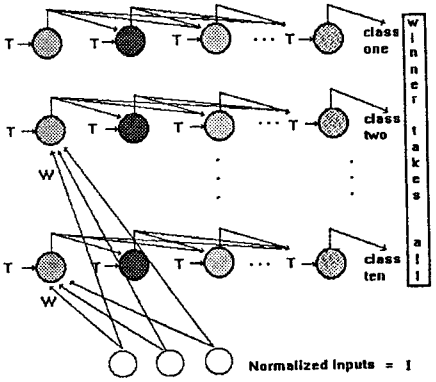

그림1.시공간 패턴 인식기의 구조

시공간 패턴 인식기는 여러 개의 층으로 구성되어 있으며 각각의 층은 서로 다른 단어에 대한 특징을 갖는 동적

뉴런들로 구성 되어 있다. 한층의 뉴런 수는 입력의 시간축 패턴갯수와 같다. 입력이 들어 오면 이 뉴런들이 동적인 활성화값을 계산하여 패턴의 일치 정도를 다음 단계의 뉴런들에 전달한다. 이때 각층의 마지막 뉴런의 최종 출력값이 가장 큰층이 winner가 된다. 각 층의 구조는 동일 하며 단지 기억 되어 있는 패턴만이 다를 뿐이다. 다음은 한 층의 동작 원리 이다.

i번째 뉴런의 총입력 (Q_i) 은 다음과 같은식을 갖는다.

$$Q_i = I \cdot W_i + d\sum_{j=1}^{L} x_k \qquad (1)$$

여기서 I는 입력 벡터이며 x_i는 다른 뉴런의 출력값이다. 그때의 출력은 다음과 같다.

$$\Delta x_i = A(-ax_i + b[Q_i - \Gamma]^+) \qquad (2)$$

위에서 함수 A(u)는 attack function이라고 부른다. 이것을 수식으로 정의 하면 다음과 같다.

$$A(u) = \begin{cases} u & \text{if } u>0 \\ cu & \text{if } u\le 0 \end{cases} \qquad (3)$$

여기서 0<c<1 이다. 이 attack function은 각 뉴런의 출력의 상승과 하강 시간에 영향을 주는 함수로써 하나의 시간적 패턴이 일치 하였을 경우에 출력이 증가 하며 다음에 오는 시간적 패턴이 비록 불일치 하더라도 출력이 급격히 감소 하지 않도록 하여 일부 노이즈가 섞이거나 발음이 다소 변화 하더라도 전반적으로 일치 하면 이를 인식하도록 하는 역할을 한다. 이 함수에 의한 시공간 패턴 인식기의 뉴런의 출력은 그림 2와 같다.

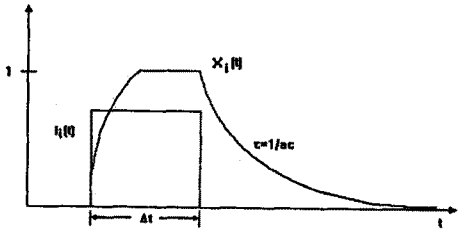

그림 2 attack function의 작용

시공간 패턴인식기에서는 여러 가지 파라미터들이 사용된다. 본논문에서는 시간의 변화에 따른 공간 패턴의 일치도와 인식률등을 기준으로 파라미터를 적절히 조절 하여 인식률을 향상 시켰다. 식 (2)에서 a 는 다음 출력에 미치는 과거 출력의 감쇄율이다. b 는 새로 들어오는 입력에 대한 가중치이다. c 는 attack function의 감쇄율로 각 뉴런의 출력량을 조절하게 된다. d 는 전단계 뉴런으로 부터 전달되는 출력량의 크기를 결정 하게 된다. Γ 는 임계치로써 패턴이 일치 되지 않는 출력에 대하여 억제 하는 효과를 갖는다.

2.3 학습 과정

시공간 패턴 인식기의 입력과 가중치 크기가 1인 벡터는 정규화 되어 있으며 이를 다음과 같이 쓸 수 있다.

$$I \cdot W = \|I\| \cdot \|W\| \cos\theta = \cos\theta \qquad (4)$$

이와 같이 벡터가 정규화 되었을때 가중치 변화량은 다음과 같이 된다.

$$\Delta W = \alpha(I - W) \qquad (5)$$

학습은 다음과 같은 알고리즘을 반복 함으로 수행 된다.

① 학습 할 층을 선정하고 입력 벡터를 가한다.
② ΔW를 식 (5)에 의하여 계산하고 이 값을 가지고 가중치를 변경 한다.

$$W(t+1) = W(t) + \alpha(I - W(t)) \qquad 0<\alpha<1$$

③ ①과 ②의 과정을 각층의 수만큼 수행 한다.
④ ③의 과정을 여러번 수행 한다.

여기서 적절한 학습 횟수는 정확히 알 수 없으며 실험적으로 구한다. 모든 학습 과정이 끝난후 가중치 벡터들은 각 그룹의 평균적인 위치에 도달 한다. 위의 과정은 그림 3에서 표현된다.

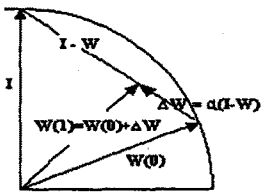

그림3. 가중치 벡터의 변경

3. 전처리 과정

음성이 마이크를 통하여 입력이 되면 전처리 보드상에서 증폭 된후 5kHz의 차단 주파수를 갖는 저역 통과 필터를 거쳐서 10 kHz의 샘플링 주파수로 A/D 변환을 한다. 이렇게 처리된 음성은 각 단어당 6개의 구간으로 나누어 해밍 윈도우를 계산한다. 그후 FFT를 통과 하여 주파수 영역으로 변환 한다. 이렇게 함으로 음성신호의 정보가 많이 손실되지 않는 범위에서 데이타량을 감소 시키고 잡음과 화자에 비교적 영향을 받지 않도록 할 수 있다. 주파수 영역으로 변환된 데이타는 24개의 채널로 나누어 각 주파수 대역의 값들을 평균하여 그 채널의 에너지 값으로 사용 하였다. 이 값들은 시공간 패턴 인식기의 입력으로 사용하기 위하여 0과 1사이의 값들로 정규화 하였다. 각 채널들은 정보량의 중요도에 따라 각구간의 크기를 mel 스케일에 의하여 지수함수 적으로 구간을 나누었다. mel 스케일에 의한 방법이 선형 적으로 공간을 분할하는 것 보다 우수한 인식률을 보였다. 입력으로 들어온 단어는 시간 프레임으로 6개의 윈도우를 갖고 공간 프레임으로 24개의 채널로 구성된 벡터의 형태로 시공간 패턴 인식기의 학습및 재현 데이타로 사용 되었다.

4. 실험 결과

이 논문에서 이용한 음성 자료는 남성 화자 2명과 여성 화자 2명이 각각 5회 발음한 숫자음 10개 (one, two, three, four, five, six, seven, eight, nine, ten)로 구성되어 있다. 이 데이타중 2회는 학습에 사용 하고 3회는 테스트용으로 사용 하였다. 이 실험에서는 패턴이 일치하지 않았을 경우 억제 시키는 파라미터 gamma (Γ)에 대하여 이를 변화 하였을 때의 인식률을 알아 보았다.

gamma	0.2	0.4	0.5	0.6	0.7
인식률	72.5%	74.2%	78.3%	80%	75.6%

실험 결과를 보았을 때 파라미터 Γ가 0.6일 때 가장 좋은 인식률을 얻었다. 하지만 실험 내용을 자세히 살펴 보면 전체적으로는 Γ가 0.6일 때 가장 인식률이 좋지만 "two"와 같은 단어에서는 Γ가 0.4일 때 오히려 더 좋은 인식률을 얻었다. 각각의 단어에 대하여 최적의 Γ값이 있으므로 입력 단어에 따라 적절히 Γ를 변경 시켜주는 알고리즘이 개발 되어야 할 것이다. 모든 학습이 끝난후 음성 데이타를 가지고 실험을 하였을 때 다음과 같은 인식 결과를 얻었다.

	1	2	3	4	5	6	7	8	9	10
1	11								1	
2		8	4							
3			10			2				
4				9	3					
5					12					
6			3			9				
7							10			2
8							1	9	2	
9	2						1		8	
10			1						1	10

위 결과를 살펴 보면 two와 eight이 저조 하게 나타 났다. 그러나 학습 데이타 수를 적절히 증가시키고 파라미터를 조절하면 인식률이 상당 부분 개선 될 것으로 보인다. 모든 음성 데이타를 가지고 인식 실험을 했을 때 최종 출력은 80%의 인식률을 얻었다.

5. 인식 과정

시공간 패턴인식기가 단어를 인식하는 과정을 살펴 보기로 하자. 입력 단어의 윈도우들이 시간에 따라 들어 올 때 시공간 패턴 인식기의 각층의 출력값은 패턴의 일치 여부에 따라 달라 진다. 뉴런의 출력은 모두 0에서 부터 시작되는데 attack function의 영향으로 전단계의 뉴런에서 패턴이 일치 되면 다음 단계에 큰 값을 계속 넘겨 줌으로 전체적으로 일치도가 가장 큰 단어를 찾아 낸다. 다음 그림4 에서는 "five" 라는 단어를 가지고 각 단어들의 출력값의 변화를 보았다. 세로축은 뉴런의 출력값을 나타낸 것이고 가로축은 각 윈도우 별로 그때 입력 되어 지는 음소를 시간에 따라 표시한 것이다. 시간 축으로 6개의 윈도우가 처리 됨으로 6개의 구간으로 나누어 표시 하였다. 그림4 에서 보는 것과 같이 "five"에 해당하는 출력값이 계속 증가함을 알 수 있다. 인식과정이 진행 되어 나가면서 유사한 음소를 가지는 단어들의 출력값도 각각 커지지만 전체적으로 일치도가 가장큰 "five"가 결국엔 출력 된다.

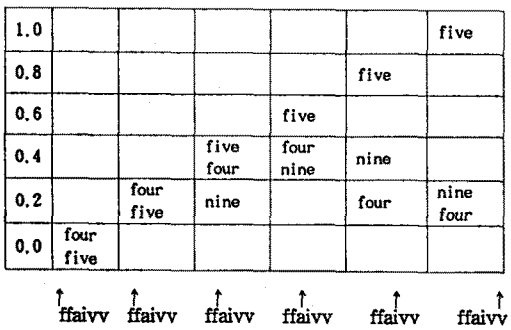

그림 4. 단어 "five"인식 하는 과정

단어 "five"를 국제 음성학회에서 정한 IPA(International Phonetic Alphabet : 국제 음성 기호)를 사용 하여 표시 하면 /faiv/가 된다.[8] 이 음소가 6개의 윈도우에 나뉘어 들어가 있다. 맨위의 단어는 현재 가장 일치하고 있는 것을 의미하며 그 밑의 단어들은 경쟁 하고 있는 것을 의미한다. 그림의 '↑' 표시는 현재 입력 되고 있는 음소를 의미 한다.

6. 결론

본 연구에서는 음성의 주파수 변동과 시간변동의 특징을 동시에 다룰 수 있는 모델로 시공간 패턴인식기를 제안 하였다. 실험은 다수의 화자에 의한 10 개의 숫자음 인식에 관하여 해보았다. 이 모델에 의한 인식률은 80%로 나타났다. 이 연구에서 가장 어려웠던 점은 시공간 패턴인식기는 특히 파라미터가 많이 있어 이 값들을 적절히 조정 하는 것이었다. 앞으로의 연구과제는 이러한 파라미터를 자동으로 조정 하는 방법을 개발 해야 할 것 이다. 시공간 패턴 인식기는 비록 메모리에 대한 손실은 다소 크지만 계산이 간단하여 처리속도가 빠르고 모듈의 구조를 갖고 있으므로 하드웨어로 구성 하기가 용이하여 대용량의 시스템을 구성 할 수 있다. 이 모델은 신경회로망의 고유한 특징인 잡음 극복 능력과 연상 기억능력을 갖고 있으면서도 신경회로망의 단점인 규모가 커질수록 학습 시간이 길어지고 학습 데이타가 폭주하는 하는 문제점을 해결할수있다.

7.참고 문헌

[1] Z.Huang and A.kuh,"A Combined Self_Organizing Feauture Map and MLP for Isolated word Recognition", IEEE Trans.signal processing,vol40,pp 2651-2657,1992.

[2] Don R.Hush and Bill G.Horne, "Progressing Supervised Neural Networks",IEEE magazine of signal processing, pp8-39,1993.

[3] Peter Ladefoged, A Course in Phonetics , Harcourt Brace Jovanovich, pp119-139,1982.

[4] David P.Morgan and Christopher L.Scofield, Neural Networks and Speech Processing,K.A.P, pp 9-39

[5] 김현기,박경철,이종호,"시공간 패턴 인식기를 이용한 숫자음 인식",인공지능,신경망및 퍼지 관련 학술 대회 논문집, pp 122 - 127,1993

[6] James.A Freeman and M.Skapura,Neural Networks ,AWPC, pp 341-371,1990

[7] Robert Hetch-Nielson,Neuro Computing,AWPC, 1990

[8]. 오 영환,패턴인식론,정익사,pp.171-172,1991