한국어 음소 인식을 위한 신경회로망에 관한 연구

최 영배*, 양 진우*, 김 순협*
*광운대학교 전자계산기공학과

A Study on Neural Networks for Korean Phoneme Recognition

Young-Bae Choi , Jin-Woo Yang , Soon-Hyob Kim
Computer Engineering Department , Kwang Woon University

ABSTRACT

This paper presents a study on Neural Networks for Phoneme Recognition and performs phoneme recognition using TDNN(Time Delay Neural Network). Also, this paper proposes new training algorithm for speech recognition using neural nets that proper to large scale TDNN.

Beause phoneme recognition is indispensable for continuous speech recognition, this paper uses TDNN to get accurate recognition result of phoneme. And this paper proposes new training algorithm that can converge TDNN to optimal state regardless of the number of phoneme to be recognized.

The result of recognition on three phoneme classes shows recognition rate of 98.1%. And this paper proves that proposed algorithm is a efficient method for high perfomance and reducing convergence time.

I. 서 론

음성 인식은 이미 그 역사가 50년에 가까운 학문 분야이다. 음성 인식은 새로운 인식 알고리즘의 개발, 고속 프로세서의 출현과 새로운 디지탈 처리 기법의 개발로 인하여 많은 발전을 거듭하여 왔다. 그러나 이러한 진전에도 불구하고 아직까지는 범용화할만한 인식 시스템이나 기법이 존재한다고 보기 어려운 실정이다. 즉, 음성 인식의 가장 큰 3가지 목표인 연속음성의 인식, 무한 어휘의 인식, 화자 독립등을 실현하기엔 기존의 기술에 의한 접근으로는 많은 난제를 안고 있으며, 음성인식의 실현시 얻을 수 있는 엄청난 부가가치에도 불구하고 음성인식 기술을 접목한 상품화 역시 미미한 단계이다.

이러한 기존 기술들의 한계를 극복하기 위하여 많은 새로운 기법들이 출현하고 있으며, 그 중 가장 대표적인 것이 거의 모든 공학 분야에서 최근 들어 각광을 받고 있는 신경망(Neural Network)이다. 신경망은 그 특유의 많은 장점들로 기존의 폰 노이만형 컴퓨터에서 잘 수행하지 못하는 종합적인 판단 과정이 필요한 패턴인식 분야에서 우수한 성능을 발휘한다[1][2]. 위와 같은 장점으로 인하여 현재 국내외에서 많은 신경망을 이용한 음성인식 논문이 발표되고 있다.

본 논문에서는 연속음의 인식이라는 궁극적인 목표를 위해서 필수적인 음소의 인식을 신경망을 이용하여 수행하였다. 연속음의 인식은 단독음의 인식과는 달리 인식 어휘의 수가 크게 증가하게되므로 부단어(sub-word)단위의 인식이 필수적이며, 이러한 부단어 단위로 음소가 가장 널리 사용되고 있다. 입력된 발성음성을 각 음소 단위로 나누고 각 음소를 인식한 후 얻은 출력 음소열로부터 발음사전의 각 단어에 대한 음소열에 대해 단어발생확률을 계산하여 최고의 확률을 내는 단어를 인식어로 선정하게되는 과정을 통해 인식을 수행하게 된다. 본 논문에서는 이러한 최종의 연속 음성 인식을 위한 한국어 음소의 인식에 관하여 연구하였으며, 아울러 신경회로망을 이용한 음성 인식의 성능 향상에 가장 큰 걸림돌이었던 기존의 학습방법인 에러 역전파(Error Backpropagation) 알고리즘을 대신할 새로운 학습 알고리즘을 제안하였다.

II. 신경회로망을 이용한 음성 인식

1. TDNN의 구조

시간지연 신경망(TDNN)은 음성의 시간축에서 연속되는 프레임들의 특징을 시간축상의 위치에 무관하게 그 특징을 추출하도록 시간지연요소 D를 둔 다층 신경망(multi-layer neural network)이다. 시간 지연요소에 의해 시간 지연이 없는 입력과 시간지연된 입력을 묶어서 네트워크에 입력으로 인가시키고 입력과 weight가 곱해진 합을 시그모이드(sigmoid) 함수와 같은 활성화(activation) 함수를 사용하여 처리한 값을 다음 층(layer)에 인가해 주는 구조로 되어 있다.

본 연구에서 사용되는 시간지연 신경망(TDNN)의 전체적인 구조는 그림 1과 같다. 기본적인 구조는 4 layer의 MLP

(Multi Layer Perceptron)의 형태를 하고 있으나, 각 layer사이의 weight의 연결에 제약을 두어 시간 지연소자를 둔 구조와 결합하여 네트워크가 탐지하려고 하는 spectral event, 즉 음소인식 신경망에서의 각 음소의 특징을 시간상의 위치에 무관하게(time shift invariance) 인식을 수행할 수 있다.

그림 1의 구조에서 음성은 각 프레임당 16 차의 특징벡터로 표현되어 네트워크의 입력으로 인가된다. 이러한 15 프레임을 음소 인식 시간지연 신경망의 입력으로 사용하여 연결이 제약되어 있는 weight에 의해 상위 층(layer)로 갈수록 더 넓은 시간대의 보다 복잡한 특성을 집약하여 전체 입력중에서 추출하고자 하는 음소의 특성을 추출하는 방식으로 네트워크가 구성된다. 본 실험에서는 16차의 특징벡터를 추출함에 있어 인간의 청각 기관의 저주파에 민감한 특징을 반영한 mel-scaled 된 파라메터를 사용하였다.

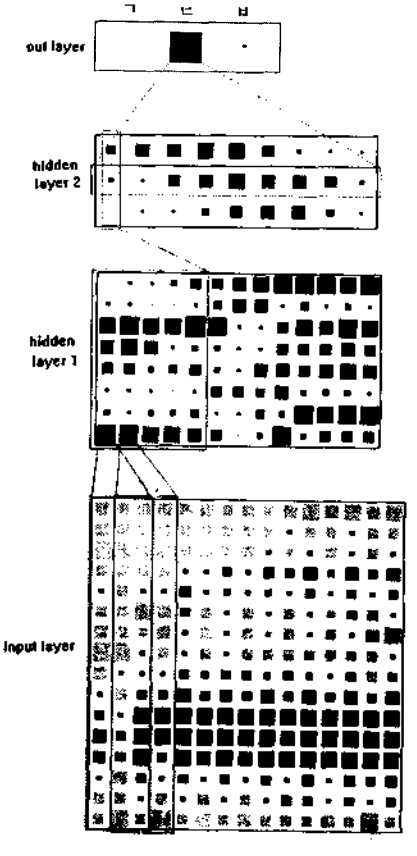

그림 1. TDNN의 구조

2. TDNN의 학습

시간지연 신경망의 학습은 기존의 MLP와 같은 에러 역전파(error Backpropagation) 알고리즘이 대부분 사용되고 있으며, 학습시간이나 인식률의 향상을 위하여 역전파 알고리즘을 조금 변형한 알고리즘들도 사용되고 있다. 그러나 시간지연 신경망은 MLP와는 달리 서로 다른 시간 간격에서 활성도를 계산하는 weight의 연결이 시간축상으로 제약되어 있는 구조를 하고 있으므로 역전파 알고리즘의 수정을 요한다. 또한 입력 층의 weight는 인식하고자 하는 음소의 특징만을 감지하려고 하므로 시간지연된 13 개의 시간대에 모두 동일한 weight를 가져야 하므로 학습이 끝날때마다 각각의 평균값으로 모두 같은 값을 갖게 된다.

학습 알고리즘의 근본 원리는 전체의 네트워크의 error값을 작게 하기 위해 기울기가 작은(negative) 방향으로 전체 weight를 조정하게 된다. 또한 뉴럴 네트워크의 가장 큰 해결 과제인 과도한 학습시간을 줄이기 위하여 음성 입력치의 처리나 weight값의 변화량등을 학습정도에 따라 수정을 하는 방법을 채택하였고, 기존의 역전파 알고리즘의 치명적인 단점인 국부적인 최소값(local minima)문제를 해결하기 위한 새로운 학습 방법을 본 논문에서 제안하였다.

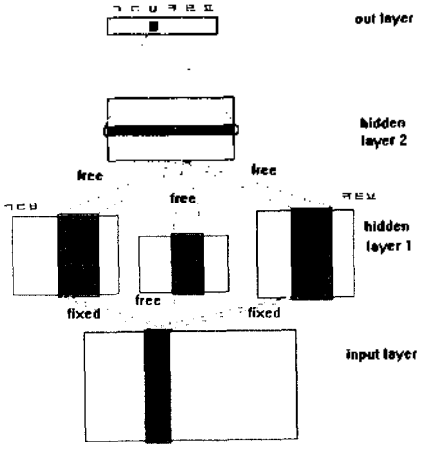

그림 2. 모듈화된 시간지연 신경 회로망

본 연구에서는 70 Hz - 4.5 KHz의 필터와 10KHz의 샘플링으로 얻은 데이타에서 인간의 청각 기관의 특징과 유사하게 음성의 특징을 추출하는 성질을 지닌 16차의 mel-scaled spectral 계수를 구하여 실험을 수행하였다.

Ⅲ. TDNN의 제안된 학습 방법

음성인식에 사용되는 신경회로망을 이용한 시스템의 구축시 가장 문제시 되는 것이 막대한 량의 학습 시간과 에러값을 최소로 하는 최적 상태로의 수렴 문제였다. 이는 대부분의 신경회로망에서 사용하고 있는 역전파(Backpropagation) 알고리즘 자체의 구조에서 기인한 문제로서 다소간의 향상된 결과를 얻기 위한 수정된 알고리즘들이 발표가 되었지만 아직도

신경회로망을 이용한 시스템 구축에 있어 적당한 학습알고리즘의 출현이 가장 선행되어야 할 문제점임이 지적되고 있다.

1. 기존 학습법의 문제점

음성 인식을 위한 신경망중 가장 우수한 성능을 보이고 있는 시간지연 신경 회로망은 다층구조로 인하여 역전파 알고리즘을 사용하여 weight를 수정하여 에러값을 줄여나가는 방식으로 학습을 수행한다.

어떠한 패턴 P에 대한 네트워크의 출력과 네트워크의 목표(Target)값과의 차에서 얻어지는 에러함수를 식 (1)과 같이 정의한다.

$$E_p = \frac{1}{2}\sum_j (t_{pj} - O_{pj})^2 \qquad (1)$$

식 (1)과 같이 정의된 에러함수 E_p를 최소로 만들도록 각 유니트간의 연결값(weight)를 변화 시키는 것이 역전파 알고리즘의 주된 원리이며, 다음과 같은 과정을 통하여 웨이트를 변화시켜 주어진 작업에 맞도록 네트워크를 최적화시키게 된다.

각 유니트의 활성도(activation)는 유니트 j에 대해 식 (2)와 같이 정의 되고, 유니트의 출력값은 유니트의 활성도를 squashing 함수를 통하여 얻어지게 되며, squashing 함수로는 시그모이드 함수(sigmoid function)가 가장 많이 사용된다.

$$net_{pj} = \sum_i \omega_{ij} O_{pi} \qquad (2)$$

$$O_{pj} = f_j(net_{pj}) \qquad (3)$$

$$f(net) = \frac{1}{1+e^{-\varepsilon \cdot net}} \qquad (4)$$

윗 식과 같이 처음에는 네트워크의 하부층에서 상부의 출력층까지의 단계가 반복되어 수행이 되고, 출력층에서 에러함수를 계산하게 되어 그 에러값의 크기에따라 에러를 줄이기 위한 연결강도의 조정인 역전파(Backpropagation)가 이루어지게 된다.

그러나, 에러 역전파 알고리즘이 항상 에러 함수의 최소값에 수렴을 하는 것은 아니다. 즉, 역전파 알고리즘은 gradient descent 형태의 알고리즘이므로 에러 표면의 기울기의 아래쪽으로만 상태가 천이하게 된다. 다차원의 에러 표면은 상당한 굴곡을 지닌 고차원의 공간과 같으므로 네트워크의 상태는 전체적인 최소점이 아닌 국부적인 최소점(Local Minima)에 빠질 수 있다는 것이다. 이러한 점은 과도한 학습시간과 함께 역전파 알고리즘을 학습알고리즘으로 채택한 많은 신경망들에게 커다란 제약을 주고 있다.

2. 새로운 학습 알고리즘의 도입

앞절에서 살펴본 바와 같이 역전파 알고리즘에서 얻을수 있는 에러 함수의 국소점이 아닌 최소점을 얻을 수 있는 새로운 알고리즘의 출현이 요구된다. 본 절에서는 에러 함수의 최소점에 수렴을 하도록 하는 코우시 학습(Cauchy Training) 알고리즘을 시간지연 신경망의 다층구조에 접목한 보다 우수한 새로운 학습법을 제안한다.

코우시 알고리즘은 본래 볼쯔만 머신(Boltzmann machine)의 수렴속도를 고속화하기 위해 사용된 알고리즘이다. 볼쯔만 머신은 단층의 상호 결합 구조의 신경망인 홉필드 네트(Hopfield Nets)가 최소점에 수렴하지 못하고 국소점에 수렴하여 국소적인 최소값(local minima)에 빠지는 것을 개량하기 위하여 동작규칙을 확률적으로 확장된 모델이다. 아래의 그림 3는 홉필드 네트가 에러 함수의 국소점, 즉 국부적인 최소값에서 전체적인 최소점으로 수렴하지 못하는 것을 개념적으로 표현하기 위하여 단순하게 2차원적으로 표현한 그림이다.

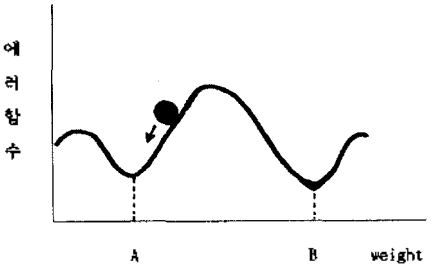

그림 3. 국부적 최소점의 문제

홉필드 네트에서는 동작원리가 네트워크의 상태가 에너지를 반드시 감소시키도록 변화하였지만, 볼쯔만 머신에서는 에너지가 증가하는 상태로의 천이도 작은 확률로 허용하는 동작규칙을 적용하여 국소적인 최소값에서 탈출하여 에너지 함수의 국소값이 아닌 최소값으로의 수렴이 가능하도록 해준다.

볼쯔만 머신의 동작은 높은 온도로 가열된 금속과 같이 네트워크의 상태를 높은 에너지 상태로 시작하여 시간 발전시켜 평형상태에 도달한 후에 평형상태를 붕괴시키지 않도록 서서히 온도를 낮추어 최종적으로 온도 0의 극한(에너지의 최소값)으로 도달하게 하는 금속의 재련시 사용되는 설담금법과 비유(Simmulated Annealing)된다.

코우시 학습법은 이러한 볼쯔만 머신의 에너지 함수의 최소값으로의 수렴하는 장점을 그대로 수용하면서, 볼쯔만 분포에 의한 많은 학습 시간을 크게 줄여 주기 위해 상태 천이 확률을 적당히 선택함으로서 코우시의 평형 분포를 사용하는 알고리즘이다.

볼쯔만 머신 :

$$P(w) = \exp(-w^2/T^2) \qquad (5)$$

$$T(t) = T_0/\log(1+t) \qquad (6)$$

코우쉬 알고리즘 :

$$T(t) = T_0/(1+t) \quad (7)$$

$$P(x) = T(t)/[T(t)^2+x^2] \quad (8)$$

$P(w)$: Size w의 연결강도 변화 확률.

$T(t)$: 인공적인 온도, 즉, 네트워크의 에러 함수의 현상태 t에서의 크기.

T_0 : 초기 Artificial Temporature.

코우시 알고리즘을 사용함으로서 학습속도를 $T(t) = T_0/\log(1+t)$에서 $T(t) = T_0/(1+t)$로 거의 극적으로 감소시킬수 있도록 해준다. 즉, 코우시 알고리즘으로는 불쯔만 머신의 최소점으로 반드시 학습하는 장점과 함께 빠른 수렴 속도도 보장 받게 된다.

코우시/역전파 결합 알고리즘의 연결 강도 조정은 두 요소를 갖는다. 그중 하나는 역전파 알고리즘을 이용해 계산된 directed 요소와 코우시 분포에 의해 결정된 random 요소이다.

$$x_c = P\{T(t)\tan[P(x)]\} \quad (9)$$

$$\omega_{mn,k}(n+1) = \omega_{mn,k}(n)+\eta[\alpha\delta\omega_{mn,k}(n)+(1-\alpha)\delta_{n,k}OUT_{m,j}]+(1-\eta)x_c \quad (10)$$

즉, 식 (10)와 같이 표현되며, 이때 η는 연결 강도의조정에 대한 코우시와 역전파 알고리즘간의 상대적인 크기를 조정하는 계수로 η가 0으로 되면 시스템은 순수한 코우쉬 머신이 되고 η가 1에 가까와지면 역전파 시스템이 된다.

이러한 결합된 새로운 학습 알고리즘으로 필기체 한자 인식을 비롯한 여러 실험에서 여타의 역전파 알고리즘이 사용된 시스템에 비해 인식률이나 수렴시간에서 우수한 성능이 보고되었다[4]. 코우쉬나 역전파중 어느 하나만을 사용한 시스템보다 두 기법을 결합한 모델이 가장 우수한 결과를 보였음이 보고되었으며, 이상의 결과로 음성 인식 분야의 신경망에 아직 적용되지 않았던 두 기법의 결합된 형태의 모델을 사용하여, 신경망의 우수한 장점들에도 불구하고 음성 인식에 폭넓게 적용되지 못한 가장 큰 결점이었던 과도한 학습시간과 국부적 최소값으로의 수렴 문제를 해결하는 새로운 방안을 본 논문은 제시한다.

Ⅳ. 실험 결과

본 논문에서는 4가지 실험을 수행하였으며, (실험 1)은 학습율을 0.01로 모멘트는 0.1로 고정하였으며, 실험 2의 모음 인식과 실험 3은 에러값과 그 변화율에 따라 학습율과 모멘트를 변화시켰다.(모멘트는 0.01 - 0.08 사이에서 변화시켰다.)

인식에 사용된 데이타의 화자의 수는 2 인이며, 실험대상 어휘는 다음과 같다.

실험대상 어휘

--- { ㄱ,ㄷ,ㅂ } * {아,이,우,에,오}
15개 음성을 5번 반복(× 2인) (= 150 개 어휘)
┌ 15 * 3번 *2 (= 90 개 어휘) ----> 학습 데이타
└ 15 * 2번 *2 (= 60 개 어휘) ----> 테스트 데이타

--- { ㅋ,ㅌ,ㅍ } * {아,이,우,에,오}
15개 음성을 5번 반복(× 2인) (= 150 개 어휘)
┌ 15 * 3번 *2 (= 90 개 어휘) ----> 학습 데이타
└ 15 * 2번 *2 (= 60 개 어휘) ----> 테스트 데이타

--- { 아, 이, 우, 에, 오 } * 3번 * 2인 (= 30 개)
{ ㄱ,ㄷ,ㅂ } * {아,이,우,에,오} * 3번 *2인 (= 90 개)
120 개의 학습데이타 (30개 + 90 개)
{ 아, 이, 우, 에, 오 } * 2번*2인 (= 20 개)
──> 테스트 데이타로 사용

1. 신경망 입력 처리 비교 실험 (실험 1)

(실험 1)에서는 화자 1인에 대하여 mel-scaled된 입력과 처리를 해 주지 않은 입력으로 실험을 진행하였다.

	처리 않된 입력	mel-scaled된 입력
ㄱ	(59/60) = 98.33 %	(60/60) = 100 %
ㄷ	(59/60) = 98.33 %	(59/60) = 98.33 %
ㅂ	(56/60) = 93.33 %	(59/60) = 98.33 %
전 체	96.67 %	98.89 %

표 1. 신경망 입력처리 비교 인식결과

2. 3개 음소군에 대한 인식 실험 (실험 2)

실험 2에서는 3개의 음소군 {ㄱ,ㄷ,ㅂ}와 { ㅋ,ㅌ,ㅍ},{ 아,이,우,에,오}에 대하여 인식을 수행하였다. 대상 음소군의 선정은 전체 음소에 대한 실험은 현실적으로 어려우므로, 크게 자음과 모음, 그리고 자음은 다시 유성자음{ㄱ,ㄷ,ㅂ}와 무성자음{ㅋ,ㅌ,ㅍ}로 나누어 실험을 수행하였다.

음 소	인식 결과
ㅋ	(60/60) = 100 %
ㅌ	(59/60) = 98.33 %
ㅍ	(58/60) = 96.67 %
전 체	98.33 %

표 2. { ㅋ,ㅌ,ㅍ } 의 인식 결과

음 소	인식 결과
아	(20/20) = 100 %
이	(19/20) = 95.00 %
우	(18/20) = 90.00 %
에	(19/20) = 95.00 %
오	(20/20) = 100 %
전 체	96.00 %

표 3. { 아,이,우,에,오 }

3. 제안된 학습알고리즘에 의한 실험(실험 3)

코우시분포를 도입한 수정된 학습알고리즘과 기존의 역전파 알고리즘과의 인식 결과는 다음 표 4와 같다.

	기존의 학습 방법	제안된 학습 방법
ㄱ	(59/60) = 98.33 %	(60/60) = 100 %
ㄷ	(58/60) = 96.67 %	(59/60) = 98.33 %
ㅂ	(56/60) = 93.33 %	(59/60) = 98.33 %
ㅋ	(60/60) = 100 %	(59/60) = 98.33 %
ㅌ	(59/60) = 98.33 %	(59/60) = 98.33 %
ㅍ	(58/60) = 96.67 %	(58/60) = 96.67 %
전 체	97.22 %	98.33 %

표 4. 제안된 학습 알고리즘의 인식결과

V. 결론

본 논문은 음성 인식의 최대의 과제중의 하나인 연속음성의 인식을 위해 반드시 선결되어야 하는 음소인식을 수행하였다. 3개 군의 음소에 대하여 인식 실험을 수행하여 {ㄱ,ㄷ,ㅂ}에 대해서는 98.89%, {ㅋ,ㅌ,ㅍ}에 대해서 98.33%, {아,이,우,에,오}에 대해서는 평균 96%의 인식률을 얻었다.

또한 시간지연 신경회로망의 광범위한 음성인식에의 적용에 걸림돌이 되고있는 기존의 에러 역전파 알고리즘의 단점을 보완하는 새로운 학습알고리즘을 제시하였다. 확률적인 방법을 도입한 코우쉬 알고리즘을 역전파 알고리즘과 결합하여 사용함으로서 보다 최적의 상태로 수렴하여 향상된 인식률과 수렴시간을 감소시킬 수 있었으며, 보다 큰 규모의 모듈화된 신경망에 적용할 경우 더욱 우수한 결과가 예상된다.

참 고 문 헌

[1] D.E. Rumelhart and J. L. McClelland, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. I and II. Cambridge, MA: M.I.T. Press, 1986.

[2] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme Recognition using time-delay neural networks," IEEE Tranx. Acoust., Speech, Signal Procesing, vol. 37, pp. 328-339, Mar.1989.

[3] P.D. Wasserman, "Combined backpropagation/Cauchy machine Neral Networks," Abstracts of the First INNS Meeting, Boston 1988, Vol. 1, pp.556 Elmsford, NY: Pergamon Press

[4] L. R. Rabiner and R. W. Schafer, " Digital Processing of Speech Signals", Prentice-hall, New Jersey, 1975.

[5] W. Ma, D. Van Comprenolle, "TDNN Labeling for a HMM Recognizer," ICASSP 90, pp.421-423, 1990.

[6] Douglas O'Shaughnessy, "Speech Communications-Human and Machine." Addison Wesley ,1987.

[7] Shafer,R.W. and Rabiner,L.R. (1975)," Digital representaton of speech signals", Proc. of IEEE 1963(4), pp.662-667

[8] Y.Komori,"Time State Neural Network for Phoneme Identification by Considering Temporal Structure of Phonemic Features", Proc. of ICASSP, pp.125-128, 1991.

[9] K. J. Lang, A. Waibel, "A Time-Delay Neural Network Architecture for Isolated Word Recognition",Neural Networks, Vol. 3, pp.23-43,1990