한국어 음소 인식을 위한 신경회로망에 관한 연구

A Study on the Neural Networks for Korean Phoneme Recognition

최 영 배*, 양 진 우*, 이 형 준**, 김 순 협*
(Young Bae Choi*, Jin Woo Yang*, Hyung Jun Lee**, Soon Hyob Kim*)

요 약

본 논문은 음소인식을 위한 신경회로망에 관한 연구로서, 시간 지연 신경회로망을 이용하여 음소인식을 수행하였다. 또한, 본 논문은 대규모 시간지연 신경망에도 적합한 음성 인식 신경망의 학습 방법에 제안한다.

연속 음성의 인식을 위해 반드시 선행되어야 하는 음소의 정확한 인식을 위하여 우수한 성능을 보이고 있는 시간지연 신경망을 사용하였으며, 인식대상 음소수가 증가하여도 신경망을 최적으로 수렴시킬 수 있는 시간지연 신경망의 새로운 알고리즘을 제시하였다. 확률론적 접근법인 코우쉬 알고리즘을 에러 역전파 알고리즘에 결합하는 시간지연 신경망의 새로운 학습 알고리즘을 사용한 실험이 수행되었다.

화자 2인을 대상으로 한 3분류의 음소군 인식 실험에서 98.1%의 인식률을 얻었으며, 제안된 알고리즘이 시간지연 신경망의 더욱 우수한 인식률과 수렴 시간의 단축에 효율적이었음을 보였다.

ABSTRACT

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

Because Phoneme Recognition is indispensable for continuous speech recognition, this paper uses TDNN to get accurate recognition result of phonemes. And this paper proposes new training algorithm that can converge TDNN to an optimal state regardless of the number of phonemes to be recognized. The recognition experiment was performed with new training algorithm for TDNN that combines backpropagation and Cauchy algorithm using stochastic approach.

*광운대학교 전자계산기공학과
**한림전문대학 전자통신과
접수일자 : 1993년 5월 10일

The results of the recognition experiment for three phoneme classes for two speakers show the recognition rates of 98.1%. And this paper yielded that the proposed algorithm is an efficient method for higher performance recognition and more reduced convergence time than TDNN.

I. 서 론

음성인식은 인간에게 가장 자연스러운 맨-머쉰 인터페이스(man-machine interface)를 구현하기 위한 가장 중요한 핵심 기술중의 하나이다. 음성인식에 관한 연구는 이미 1950년대에 시작되어 많은 인식 기술들이 개발되었고, 이를 사용한 인식 시스템도 개발이 되어 왔다. 그러나 기존의 음성 인식 기법들은 발성음의 앞뒤가 묵음으로 된 경우의 단독음에 대하여는 어느정도 우수한 실험결과를 얻고 있지만, 연결어나 연속음으로 보다 그 범위를 확대해갈수록 실제 응용에 적합하지 못한 낮은 인식률을 보이고 있으며, 또한 화자의 수나 특성에도 매우 제한적인 형편이다. 이에 기존의 인식 기법으로는 연속어 인식, 화자 독립, 무제한 어휘의 인식이라는 음성인식의 최종 3대 목표를 달성하기 어려운 실정이다. 이에 새로운 돌파구로서, 현재 공학의 광범위한 분야에서 새로운 가능성을 제시하여 각광받고 있는 신경회로망을 이용한 음성 인식에의 접근이 최근 국내외에서 연구되고 있다[1][2][4].

본 논문은 음성 인식에 적용된 신경 회로망의 여러 모델중 우수한 성능을 보이는 시간지연 신경회로망(Time Delay Neural Networks)을 이용하여 음성 인식을 수행하였다[2][4][8][12]. 위에서 언급한 음성 인식의 궁극적인 3가지 목표를 실현함에 있어 신경회로망을 이용한 기술을 통한 접근은 가장 실현 가능성이 높은 기법으로 평가되고 있다. 신경회로망은 외부의 입력을 학습을 통하여 그 특징만을 추출하고 이를 일반화할 수 있는 우수한 장점을 지니고 있으므로 화자간의 특성에 따른 입력의 차이에 의한 인식 오류를 배제할 수 있으며, 우수한 패턴 분류능력과 분산처리, 병렬 처리의 구조등으로 인하여 연속음인식이나 실시간 음성 인식등과 같은 위의 목적에 가장 적합한 인식 기법이라 할 수 있다[1][4].

본 논문에서는 연속음의 인식이라는 궁극적인 목표를 위해서 필수적인 음소의 인식을 신경망을 이용하여 수행하였다. 연속음의 인식은 인간의 자연스러운 발성의 인식이므로 단독음의 인식과는 달리 인식 어휘의 수가 크게 증가하게되므로 부단어(sub-word) 단위의 인식이 필수적이며, 이러한 부단어 단위로 음소가 가장 널리 사용되고 있다. 발성된 입력음성을 각 음소단위로 나누고 각 음소를 인식한 후 얻은 출력 음소열로부터 발음사전의 각 단어에 대한 음소열에 대해 단어 발생확률을 계산하여 최고의 확률을 내는 단어를 인식어로 선정하게되는 과정을 통해 연속음의 인식을 수행하게 된다.

본 논문에서는 이러한 최종의 연속 음성 인식을 위하여 시간지연 신경회로망을 이용하여 한국어 음소의 인식에 관하여 연구하였으며, 아울러 신경회로망을 이용한 음성 인식의 성능 향상에 가장 큰 걸림돌이었던 기존의 학습방법인 에러 역전파(error back-propagation) 알고리즘을 대신할 확률론적인 접근방식인 코우시 알고리즘과 결합한 시간지연 신경망의 학습 알고리즘을 제안하였다[1][10].

II. 시간지연 신경회로망

2.1 TDNN의 학습

시간지연 신경망의 학습은 기존의 MLP와 같은 에러 역전파 알고리즘이 대부분 사용되고 있으며, 학습 시간이나 인식률의 향상을 위하여 역전파 알고리즘을 조금 변형한 알고리즘들도 사용되고 있다. 그러나 시간지연 신경망은 MLP와는 달리 서로 다른 시간 간격에서 활성도를 계산하는 연결 강도의 연결이 시간축상으로 제약되어 있는 구조를 하고 있으므로 자연히 역전파 알고리즘의 수정을 요한다. 또한 입력층의 연결 강도는 인식하고자 하는 음소의 특징만을 감지하려고 하므로 시간지연된 입력층의 13개의 시간대에 모두 동일한 연결 강도를 가져야 하므로 학습이 끝날때마다 각각의 평균값으로 모두 같은 값을 갖게 된다. 이러한 음성의 시간변화를 흡수하려고 하는 구조로 인하여 입력층의 출력은 3 프레임의 입력이 인가된 후에 처음으로 첫번째 은닉층에 출력값을 보내게 되고, 첫번째 은닉층 또한 5 프레임의 정보를 집약한 후 두번째 은닉층에 출력값을 보내게 된다. 이렇게 얻어진 값들은 입력층의 15 프레임을 집약한 결과인 두번째 은닉층의 9 프레임에 걸쳐서 두번째 은닉층의 활성도를 집약하여 같은 유니트의 9 프레임에

걸쳐 모두 같은 가중치를 갖는 연결 강도에 의해 출력층에 그 값이 전달되며, 출력층에서는 이 값들과 연결 강도들의 곱을 합하여 시그모이드 함수를 통하여 출력값을 생성하며 이 값으로 출력 노드가 각각 대표하고 있는 음소중 가장 큰 값을 갖는 것이 인식된 음소(신경망의 경우 'ㄷ')로 결정되게 된다.

그림 2.1은 {ㄱ,ㄷ,ㅂ,ㅋ,ㅌ,ㅍ}의 모듈화된 신경회로망으로 {ㄱ,ㄷ,ㅂ}와 {ㅋ,ㅌ,ㅍ}의 학습된 음소 구분 신경망을 모듈화하여 학습을 시키는 신경망이다. 많은 수의 음소를 인식할 수 있는 대규모의 시간지연 신경망을 구성하기 위해서는 기존의 방식대로 출력층에 노드의 수만을 추가한 신경망을 사용하는 것이 아니라 각각의 비슷한 특징을 갖는 음소별로 그룹을 나누어 각 그룹별로 먼저 학습을 시켜 연결강도를 조정한 후 각각 학습된 소그룹들을 모듈화하여 보다 많은 수의 음소를 인식할 수 있도록 구성하는 것이 보다 효과적이다[9][10]. 본 논문에서 사용된 모듈화된 신경망을 보여주는 그림 2.1에서는 모듈화되지 않았던 시간지연 신경망과 비교할때 더욱 복잡한 구조를 하고 있다. 개별적으로 학습된 {ㄱ,ㄷ,ㅂ}와 {ㅋ,ㅌ,ㅍ}의 그룹 신경망을 중앙의 glue 네트워크와 결합하여 보다 효율적으로 두 음소군을 인식할 수 있는 신경망을 구성한다. 그림 2.1에서 fixed로 표시된 부분의 연결의 연결 강도값은 고정시키고 free로 표시된 부분의 연결 강도값은 전체의 모듈화된 신경망의 학습에 의해 결정되게 된다. 이때 그림 2.1의 중앙의 glue

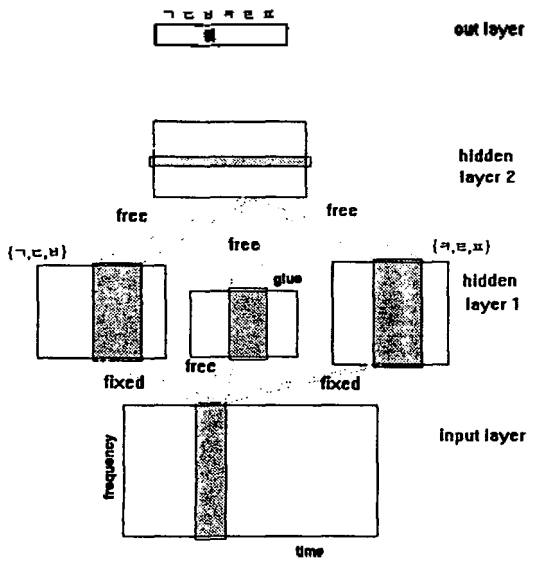

그림 2.1 모듈화된 시간지연 신경 회로망

layer는 {ㄱ,ㄷ,ㅂ}와 {ㅋ,ㅌ,ㅍ}간의 유, 무성음간의 차이를 구분하는 역할을 수행하여 각각의 학습된 두 음소분류 신경망의 기능을 통합하여 보다 큰 규모의 신경망을 효율적으로 구성, 학습할 수 있도록 해준다.

시간지연 신경망의 학습 알고리즘의 근본 원리는 전체의 네트워크의 오차 값을 작게 하기 위해 기울기가 음(negative)의 방향으로 전체 연결 강도를 조정하게 된다. 또한 뉴럴 네트워크의 가장 큰 해결과제인 과도한 학습시간을 줄이기위하여 음성 입력치의 처리나 연결 강도값의 변화량등을 학습정도에 따라 수정을 하는 새로운 방법을 채택하였고, 기존의 역전파 알고리즘의 치명적인 단점인 국부적인 최소값(local minima)문제를 해결하기 위한 시간지연 신경망을 위한 학습 방법[1][3][7]을 본 논문에서 제안하였다.

2.2 신경망 입력을 위한 음성 처리

음성은 시각적인 패턴 인식에서와는 달리 시간적인 정보가 매우 중요한 특징을 갖고 있다. 또한 음성은 이러한 시간적인 변화에 따라 계속 변화하게되므로 작은 시간구간으로 음성을 나누어 프레임별로 음성을 처리하게 되는데 본 연구에서는 10msec 프레임율로 각 프레임을 구한다. 이렇게 짧은 시간 구간동안에는 그 구간내의 음성의 특징이 어느정도 안정(quasi-stationary)된다고 볼 수 있으므로 그 구간에 있는 음성의 특징을 잘표현할 수 있도록 특징을 추출한다[5][6][13].

그림 2.2에서는 발생된 전체 음성 구간에서 세그멘테이션을 통하여 인식하고자 하는 음소가 속해있는 15프레임을 처리하여 신경망의 입력으로 사용하게 된다. 그림 2.2의 세 그림은 각각 전체 발성 음성 구간의 파형과 인식하고자 하는 음소가 위치한 구간의 파형이며, 가장 아래의 그림은 결정된 15프레임의 스펙트로그램을 표현한 것이다.

본 연구에서는 인간의 청각 기관의 특징과 유사하게 음성의 특징을 추출하는 성질을 지닌 16차의 mel-scaled spectral 계수를 구하여 실험을 수행하였다. 인간의 청각기관은 모든 주파수 대에 걸친 spectral 정보를 같은 중요 도로 취급하지 않고 저주파수대역의 정보에 많은 비중을 두고 높은 주파수 대역의 정보에는 상대적으로 작은 비중을 두는 방식으로 청각기관에 입력된 음성의 정보로 부터 특징을 추출하여 인식을 행한다는 사실로 부터 mel-scaled된 spectral

계수를 사용하게 되었으며 이는 인간의 청각적인 인식 구조를 모델링하는 것을 원칙으로 하는 신경회로망의 입력에 가장 적합하다고 할 수 있으며 mel-scaled spectral 계수는 최근의 국내외의 논문발표결과에서 음성인식 신경회로망의 특징 파라메터로 가장 우수한 인식률을 보임이 보고되었다[2][9]. 그림 2.3은 mel-scaled spectral 계수와 선형적으로 구한 spectral 계수를 비교한 그림이다.

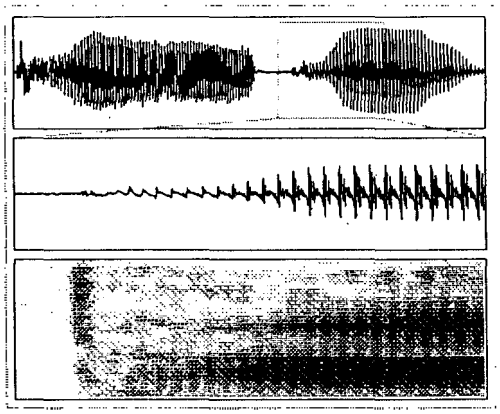

그림 2.2 신경망 입력 프레임의 결정

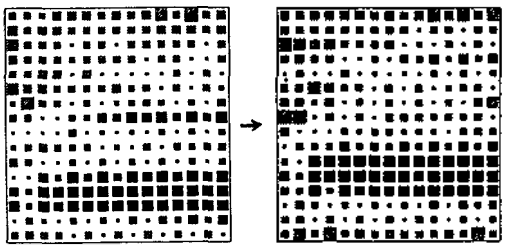

그림 2.3 mel-scaled 처리 전, 후의 입력

또한 신경망의 학습시 계산량을 줄이기 위하여 모든 입력값을 −1에서 1사이의 값을 갖도록 정규화시켜 사용하였으며[2][8], 본 논문에서 사용된 전체적인 신경망 입력을 위한 음성 처리는 그림 2.4와 같이 처리되었다.

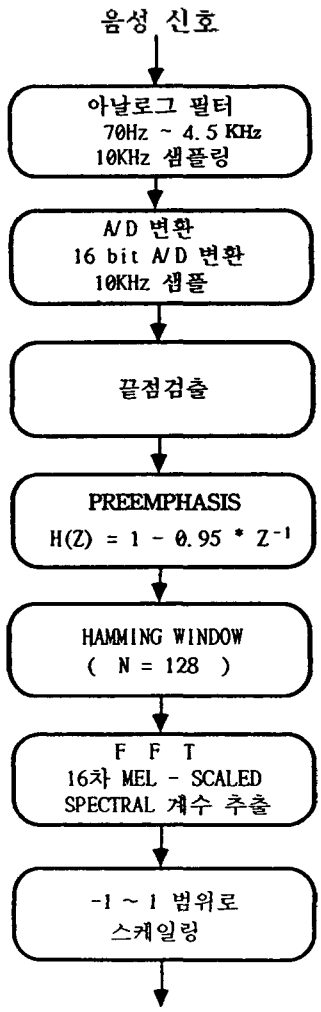

그림 2.4 음성 신호 처리의 흐름도

Ⅲ. TDNN의 제안된 학습 방법

음성인식에 사용되는 신경회로망을 이용한 시스템의 구축시 가장 문제시되는 것이 막대한 량의 학습시간과 에러값을 최소로 하는 최적 상태로의 수렴 문제였다. 이는 대부분의 신경회로망에서 사용하고 있는 역전파 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. 본 장에서는 기존의 학습법인 역전파 알고리즘에 관하여 살펴보고, 역전파 알고리즘이 가지고 있는 문제점을 해결해 줄 수 있는 학습 알고리즘을 제안한다[1][7].

3.1 기존 학습법의 문제점

음성 인식을 위한 신경망중 가장 우수한 성능을 보이고 있는 시간지연 신경회로망은 다층구조로 인하여 역전파 알고리즘을 사용하여 연결 강도를 수정하여 에러값을 줄여나가는 방식으로 학습을 수행한다. 본절에서는 기존의 시간지연 신경망의 학습방법인 에러 역전파 알고리즘에 대해 살펴보고 그 문제점에 대하여 논한다.

역전파 알고리즘은 입력층, 중간층, 출력층과 같이 다층구조를 한 네트워크에 대해서, Rumelhart등에 의해서 제안된 지도 학습(supervised learning)알고리즘이다[1][7]. 다층 구조의 네트워크를 학습시키는 우수한 성능으로 대부분의 다층 신경회로망에서 사용되고 있다.

어떠한 패턴 P에 대한 네트워크의 출력과 네트워크의 목표값과 차에서 얻어지는 에러함수를 식(3-1)과 같이 정의한다.

$$E_p = \frac{1}{2} \sum_i (t_{pj} - O_{pj})^2 \qquad (3\text{-}1)$$

식 (3-1)과 같이 정의된 에러함수 E_p를 최소로 만들도록 각 유니트간의 연결강도(weight)을 변화 시키는 것이 역전파 알고리즘의 주된 원리이며, 다음과 같은 과정을 통하여 연결강도를 변화시켜 주어진 작업에 맞도록 네트워크를 최적화시키게 된다.

각 유니트의 활성도는 유니트 j에 대해 식 (3-2)와 같이 정의 되고, 유니트의 출력값은 유니트의 활성도를 activation 함수를 통하여 얻어지게 되며, activation 함수로는 시그모이드 함수가 가장 많이 사용된다.

$$net_{pj} = \sum_i \omega_{ij} O_{pi} \qquad (3\text{-}2)$$

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

$$f(net) = \frac{1}{1 + e^{-k \cdot net}} \quad (0 < f(net) < 1) \qquad (3\text{-}4)$$

윗 식과 같이 처음에는 네트워크의 하부층에서 상부의 출력층까지의 단계가 반복되어 수행이 되고, 출력층에서 에러 함수를 계산하게 되어 그 에러값의 크기에따라 에러를 줄이기 위한 연결강도의 조정인 역전파가 이루어지게 된다.

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

3.2 새로운 학습 알고리즘의 도입

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

코우시 알고리즘은 본래 볼쯔만 머쉰(Boltzmann machine)의 수렴속도를 고속화하기 위해 사용된 알고리즘이다. 볼쯔만 머쉰은 단층의 상호 결합 구조의 신경망인 홉필드 네트(Hopfield Nets)가 최소점에 수렴하지 못하고 극소점에 수렴하여 국소적인 최소값(local minima)에 빠지는 것을 개량하기 위하여 확장된 모델이다.

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

볼쯔만 머쉰의 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 제련시 사용되는 설담금법과 비유(simulated annealing)된다.

코우시 학습법은 이러한 볼쯔만 머쉰의 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 머쉰과 코우쉬 알고리즘의 비교

①볼쯔만 머쉰 :

$$P(w)=\exp(-w^2/T^2) \tag{3-5}$$

$$T(t)=T_0/\log(1+t) \tag{3-6}$$

②코우쉬 알고리즘 :

$$T(t)=T_0/(1+t) \tag{3-7}$$

$$P(x)=T(t)/[T(t)^2+x^2] \tag{3-8}$$

$P(w)$: Size w의 연결강도 변화 확률.

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

T_0 : 초기 인공 온도

코우시 알고리즘을 사용함으로써 학습 속도를 $T(t)=T_0/\log(1+t)$에서 $T(t)=T_0/(1+t)$로 거의 극적으로 감소시킬수 있도록 해준다. 즉, 코우시 알고리즘으로는 볼쯔만 머쉰의 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.

$$x_c=P\{T(t)\tan[P(x)]\} \tag{3-9}$$

$$\omega_{mn,k}(n+1)=\omega_{mn,k}(n)+\eta[\alpha\Delta\omega_{mn,k}(n)+(1-\alpha)\delta_{n,k}OUT_{m,j}]+(1-\eta)x_c \tag{3-10}$$

즉, 식 (3-10)과 같이 표현되며, 이때 η는 연결 강도의 조정에 대한 코우시와 역전파 알고리즘간의 상대적인 크기를 조정하는 계수로 η가 0으로 되면 시스템은 순수한 코우쉬 머쉰이 되고 η가 1에 가까와지면 역전파 시스템이 된다.

이러한 결합된 학습 알고리즘으로 필기체 한자 인식을 비롯한 여러 실험에서 여타의 역전파 알고리즘이 사용된 시스템에 비해 인식률이나 수렴시간에서 우수한 성능이 보고되었으며[4][7], 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 및 고찰

본 장에서는 4가지의 실험을 수행하였으며, 실험 1은 학습율을 0.01로 모멘트는 0.1로 실험하였으며, 실험 2의 모음 인식과 실험 3은 에러값과 그 변화율에 따라 학습율과 모멘트를 변화시켰다[1][7]. (모멘트는 0.01−0.08 사이에서 변화시켰다.)

4.1 신경망 입력 처리 비교 실험 (실험 1)

(실험 1)에서는 화자 2인에 대하여 mel-scaled된 입력과 처리를 해 주지 않은 입력으로 실험을 진행하였다. 화자는 남성 2인이며 대상음소는 {ㄱ, ㄷ, ㅂ}를 선정하였다.

실험대상 어휘 : {ㄱ, ㄷ, ㅂ} × {아, 이, 우, 에, 오}
15개의 어휘를 5번 반복(×2인)
(=150개 어휘)

15×3번×2=90개의 학습 데이타(각 음소별 30개)
15×2번×2번=60개의 인식 테스트 데이타
(각 음소별 20개)

그림 4.1. 신경망 입력 처리 비교 인식 결과

	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 4.1의 결과로 mel-scaled된 입력이 보다 우수한 성능을 나타내었으며, 인간의 청각이 고주파 대역보다는 저주파 대역에 더 민감하므로 저주파대역에 가중치를 두고 그 이상의 대역의 로그함수 처리로 고주파 대역으로 갈수록 가중치를 작게 둔 입력을 TDNN에 입력하여 더 우수한 성능을 보였다[2][8].

4.2 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 실험이 현실적으로 어려우므로, 크게 자음과 모음, 그리고 자음은 다시 유성자음{ㄱ, ㄷ, ㅂ}와 무성자음{ㅋ, ㅌ, ㅍ}로 나누어 실험을 수행하였으며 각각의 실험 결과는 표 4.1, 표 4.2, 표 4.3과 같다.

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

실험대상 어휘

①{ㄱ,ㄷ,ㅂ}×{아,이,우,에,오}
15개의 어휘를 5번 반복 (×2인)(=150개 어휘)
- 15×3번×2=90개의 학습 데이타 (각 음소별 30개)
- 15×2번×2번=60개의 인식 테스트 데이타 (각 음소별 20개)

②{ㅋ,ㅌ,ㅍ}×{아,이,우,에,오}
15개의 어휘를 5번 반복 (×2인)(=150개 어휘)
- 15×3번×2=90개의 학습 데이타 (각 음소별 30개)
- 15×2번×2번=60개의 인식 테스트 데이타 (각 음소별 20개)

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

그림 4.2. {ㅋ,ㅌ,ㅍ}의 인식 결과

음 소	인 식 결 과
ㅋ	(60/60)=100%
ㅌ	(59/60)=98.33%
ㅍ	(58/60)=96.67%
전 체	98.33%

그림 4.3. {아,이,우,에,오}의 인식 결과

음 소	인 식 결 과
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%

4.3 시간축 변화 흡수 실험 (실험 3)

본 실험에서는 시간지연 신경회로망의 가장 큰 장점인 시간축의 변화가 있는 입력에 대해서도 인식률의 저하가 없는지에 대하여 실험하였다. 음소 인식실험시 세그멘테이션에 많은 시간을 소비하게 되는데 이를 자동으로 처리한 결과와 수작업으로 세그멘테이션 한 결과를 비교하였다. 단시간 로그 에너지와 영교차율을 이용한 자동 세그멘테이션과 수작업의 세그멘테이션의 결과는 크게는 4프레임의 차이에서 보통 2 프레임 이하의 차이가 있었으나 두 세그멘테이션 방법의 비교 실험 결과로 얻은 인식률에는 큰 차이가 없으므로 구현된 시간지연 신경회로망의 신축 변화 흡수(time shift invariance)능력이 우수함을 증명하였다[2][8]. 실험 데이타는 실험 1과 동일하게 구성하였다.

그림 4.4. 시간축 변화 흡수 실험 인식 결과

	핸드 세그멘테이션	자동 세그멘테이션
ㄱ	(60/60)=100%	(60/60)=100%
ㄷ	(59/60)=98.33%	(58/60)=96.67%
ㅂ	(56/60)=98.33%	(59/60)=98.33%
전 체	98.89%	98.33%

4.4 제안된 학습알고리즘에 의한 실험 (실험 4)

본 논문에서 제안한 알고리즘과 기존의 역전파 학습알고리즘을 향상시킨 신경망과의 실험결과를 비교하였다.

그림 4.5. 제안된 학습법을 사용한 인식 결과

	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%

위의 실험에서 제안된 학습 방법을 {ㄱ,ㄷ,ㅂ,ㅋ,ㅌ,ㅍ}를 모듈화한 시간지연 신경망에 적용하여 기존의 학습알고리즘에 비교시 1.11% 높은 인식률을 얻을수 있었으며, 학습의 수렴 시간도 20%이상 단축된 결과를 얻을 수 있었다. 이러한 모듈화는 음소인식 신경망의 규모를 확대할 경우의 가장 우수한 방법으로[9][10], 본 논문에서 제안된 학습법과 결합하여 보다 많은 음소군의 향상된 인식률을 얻는데 가장 적절한 방법이라고 할 수 있다[4][11]. 위의 실험 결과로 본 논문에서 제안된 알고리즘이 신경망의 우수한 인식 성능과 보다 빠른 학습 속도를 얻는데 효과적이었음을 보였다.

4.5 고 찰

실험 1에서는 신경망의 입력벡터 추출에 관한 실험

을 통하여, mel-scaled된 입력이 더 우수한 결과를 얻을 수 있었다. 이로서 신경망의 입력으로 인간의 청각기관의 주파수대에 대한 민감도를 유사하게 처리한 입력값이 신경망의 입력으로 적합함을 실험을 통하여 확인할 수 있었다[2].

실험 2에서는 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 얻을 수 있었다. 이는 화자 2인에 대한 실험으로 화자의 수를 늘리는것이 우선 과제이며, 모음의 인식률이 상대적으로 낮은 것은 적은 량의 학습 데이타와 테스트 데이타로 인하여 정확한 인식률을 얻을 수 없었던 것으로 생각된다. 또한 3개 음소군에서 더욱 확장된 거의 모든 음소를 포함하는 실험이 앞으로의 연구과제라 할 수 있다.

실험 3에서는 시간축의 변화를 시간지연 신경망이 잘 흡수할 수 있는지 실험하였다. 핸드 세그멘테이션된 데이타로는 {ㄱ,ㄷ,ㅂ}음소군에 대해 98.89%, 자동 세그멘테이션된 데이타로는 98.33%의 인식률을 얻어 상대적으로 부정확한 자동 세그멘테이션으로 얻은 데이타도 거의 대등한 인식율을 얻었으므로 인식과정에서 수작업을 배제할 수 있는 확장성을 얻게 되었고, 구성된 네트워크의 시간 변화 흡수능력을 확인할 수 있었다[2][4].

마지막으로 실험 4에서는 본 논문에서 제안한 알고리즘과 기존의 학습알고리즘과의 비교를 통한 실험을 수행하였는데, {ㄱ,ㄷ,ㅂ,ㅋ,ㅌ,ㅍ}를 글루(glue) 네트워크를 이용하여 모듈화한 시간 지연 신경망(그림 2.3)에 적용한 실험에서 제안된 학습법으로 98.33%의 인식률로 기존의 학습법보다 1.11% 우수한 인식률과 수렴 속도면에서 20%이상 개선된 결과를 얻었다. 그러나 아직 많은 실험 결과를 통하여 새로 제안된 학습법에 사용되는 각 파라메터들의 조정에 따른 충분한 실험을 통하여 최적값을 구하고 조정하는 과정들이 추가되어야 하며 이를 통하여 더욱 향상된 인식 결과를 기대할 수 있을것으로 보인다. 또한 보다 큰 모듈화된 신경망에 사용시에도 최적으로 수렴할 수 있는 학습알고리즘으로의 개선과 수정이 필요하며 기존의 학습법에 의해 제약되었던 문제들을 상당부분 해결할 수 있을 것으로 기대된다.

V. 결 론

본 논문은 음성 인식의 최대의 과제중의 하나인 연속음성의 인식을 위해 반드시 선결되어야 하는 음소 인식을 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 얻었으며, 화자 2인을 대상으로한 3 분류의 음소군 인식 실험에서 98.1%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 사용하므로서 보다 최적의 상태로 수렴하여 향상된 인식률과 수렴시간을 감소시킬 수 있었으며 보다 큰 규모의 모듈화된 신경망에 적용할 경우 더욱 우수한 결과가 예상된다.

본 논문의 연구 결과는 보다 큰 규모의 음소군의 인식을 위한 개선과 보완을 통하여 음소단위의 연속음인식을 위해 필수적인 전체 음소의 인식을 위한 보다 큰 규모의 신경망의 구성과 그 학습을 위해 효과적으로 사용될 수 있으리라 기대되며, 화자수의 증가와 보다 우수한 인식률을 위한 노력등이 계속되어져야할 과제이다.

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 Trans. 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. D.P. Morgan, C.L. Scofield, "Neural Networks and Speech Processing," Kluwer Academic Publishers, 1991.
5. L.R. Rabiner, R.W. Shafer, "Digital Processing of Speech Signals," Englewood Cliffs, N.J., Prentice-Hall, 1978.
6. Markel, J.D. and Gray, A.H. Jr. (1976), "Linear prediction of speech," Springer-Verlag.

7. Philip D. Wasserman, "Neural Computing : Theory and Practice," ANZA Research, Inc. 1989.
8. K. J. Lang, A. Waibel, "A Time-Delay Neural Network Architecture for Isolated Word Recognition," Neural Networks, Vol. 3, pp.23-43, 1990.
9. A.Waibel, H.Sawai, K.Shikano, "Modularity and Scaling in Large Phonemic Neural Networks," IEEE Trans. of Acoustics, Speech and Signal Processing, Vol.37, No.12, Dec. 1989.
10. H.Swai, A.Waibel, P.Haffner, M.Miyatake and K. Shicano, "Parallelism, Hierachy, Scaling in Time-Delay Neural Networks for Sppoting Japanese Phonemes/CV-Syllables," Proc. of IJCNN, Vol.2, pp.81-88, Washington D.C., June 1989.
11. M.Miyatake, H.Sawai, Y.Minami and K.Shikano, "Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time-Delay Neural Networks," Proc. of ICASSP, Vol.1, pp.449-452, 1990.
12. A.Hirai, A.Waibel, "Phoneme-based Word Recognition by Neural Network-A Step toward Large Vocabulary Recognition," Proc. of IJCNN, Vol.3, pp. 671-676, San Diego, California, June 1990.
13. Shuzo Saito and Kazuo Nakata, "Fundamentals of System Signal Processing," Tokyo, 1985.

▲崔 榮 培(정회원) 1967년 5월 14일생

1991년 2월 : 광운대학교 전자계산기 공학과 졸업(공학사)
1993년 2월 : 광운대학교 대학원 전자계산기 공학과 졸업(공학석사)
1993년 1월~현재 : 대우전자 중앙연구소 HDTV 팀
※주관심분야 : Digital Signal Processing High Definition TV System, Neural Networks

▲梁 鎭 宇 : 정회원, 종신회원
제 12권 3호 참조

▲金 淳 協 : 제 12권 3호 참조

▲李 炯 俊(정회원) 1956년 12월 9일생

1980년 2월 : 아주대학교 전자공학과 졸업(공학사)
1982년 2월 : 광운대학교 대학원 전자통신과 졸업(공학석사)
1985년 3월~1988년 2월 : 광운대학교 대학원 전자계산기공학과 박사과정 수료
1983년 1월~1993년 2월 : 삼성전자 종합연구소 선임연구원
1993년 3월~현재 : 한람전문대학 전자통신과 전임강사
※주관심분야 : Digital Signal Processing(음성합성, 인식) Neural Network, ASIC