분산 신경망을 이용한 고립단어 음성에 나타난 음소 인식
Recognition of Korean Phonemes in the Spoken Isolated Words Using Distributed Neural Network

김 선 일*, 이 행 세*
(Seonil Kim*, Haing Sei Lee*)

※본 논문은 1995년도 아주대학교 교내연구비 지원에 의해 연구된 것입니다.

요 약

본 논문에서는 총 106개의 단어로 구성되는 30개의 한국어 속담 문장에 대해 프레임 단위로 인식하는 분산 신경망을 구현하였다. 음성에 대한 특징값으로는 PLP 켑스트럼과 에너지 및 영교차율을 사용하였으며 분산 신경망의 입력으로 사용되는 이 특징값들이 음성의 시간적 특성을 잘 반영할 수 있도록 한 프레임 주변의 넓은 영역에 걸쳐 데이타를 수집하였다.

20대 젊은 남자가 30개의 속담을 5번씩 발음하였다. 신경망 학습에 네집단을 사용하고 학습에 참여하지 않은 나머지 한집단은 인식용으로 사용하였다.

속담내의 단어와 단어 사이는 구별이 잘 되도록 묵음 구간을 두어 발음하였다. 인식 결과 음소를 각 군별로 분류하는 대분류 신경망에서의 각 군의 프레임 인식율은 네 집단을 학습에 사용한 경우 95.3%를 나타내었다.

Abstract

In this paper, we implemented distributed neural network that recognizes phonemes by frame unit for the 30 Korean proverbs sentences consist of 106 isolated words. The features of speech were chosen as PLP cepstrums, energy and zero crossings, where we get those being used as inputs to the distributed neural networks in wide area for a frame to get the good temperal characteristics. A young man of twenties has produced 30 proverbs 5 times. The learning of neural network uses 4 sets of them. 1 set being unused remains for test. There exists silence between words for the easy discrimination. The ratio of frame recognition in large grouping neural network is 95.3% when 4 sets were used for the learning.

I. 서 론

사람과 사람 사이의 의사 전달 방법은 여러가지가 존재하지만 가장 단순하면서도 쉽게 이용되는 방법이 음성을 통한 정보 교환 방법이다. 특별한 문자를 습득하지 않아도 태어나면서 자연스럽게 접하게되는 언어를 음성을 통하여 구사함으로서 정보의 교류가 이루어지게 된다.

인간의 가장 손쉬운 정보 교류 방법이 인간과 기계로 눈을 돌려보면 그리 쉽지 않은 문제임을 알 수 있다. 기계는 기본적으로 약속된 코드에 의해 정보의 교류가 이루어지게 되어 있어서 기계와의 대화에서는 음성을 이 코드로 변환해 주어야하는 제약이 따르게 된다. 이 과정을 우리는 음성인식이라고 한다. 음성인식의 궁극적 목표는 사람과 사람 사이의 자연스런 대화를 인간과 기계 사이에 실현하는 것이다. 그러나 선진국들의 오랜 노력에도 불구하고 아직까지 이 목표는 달성되지 못하고 있으며 앞으로도 가까운 장래에 실현되지는 않을 전망이다[1]. 따라서 고립단어나[6] 숫자의 인식[7]등 제한된 영역에서 사용하고자 하는 노력들이 이루어지고 있다.

음성 인식 방법에는 여러가지가 있지만 가장 널리 쓰이는 방법으로는 DTW 및 HMM에 의한 template matching, 최근 들어 각광 받는 인공 신경망에 의한 방법등이 있다[2].

인간의 청각은 주파수의 변화에 따라 소리를 인지하며 각 주파수 영역에서의 인지도가 서로 다르다. 기계에 음성을 인식시키고자 할 때 사람과 유사한 형태로 시도하는 것이 가장 바람직한데 사람의 청각 신경을 모방한 인지 선형예측법(PLP : Perceptual Linear Prediction)[3, 4, 5]을 사용한 PLP 켑스트럼을 이용하여 해당 프레임 주변

*아주대학교 전자공학과
Dept. of Elec. Eng., Ajou Univ.
접수일자 : 1995년 9월 18일

170ms에 대해 특징값을 구하여 이를 신경망의 입력으로 사용하였다.

해당 프레임 주변의 7개 구간에 대해서 7차 PLP 및 영교차율, 에너지, 총 9개의 특징값을 사용하므로 신경망의 입력단은 63개가 되고 출력으로는 34개의 음소를 사용하게 되므로 상당히 많은 연결 상태가 존재하고 이를 이용하여 학습하려면 학습에 상당한 시간이 소요되게 된다. 따라서 입력된 프레임을 비슷한 음소군으로 대분류하고 분류된 결과를 이용하여 해당 군의 신경망으로 학습을 하면 학습 효율을 높일 수 있다. 이를 분산 신경망이라고 하면 분산 신경망은 대규모 학습에 있어 그 효능을 발휘할 수 있다.

분산 신경망은 각 프레임을 음소별로 인식하고 인식된 프레임 열은 적절한 후처리 과정을 거쳐 음소열로 바뀌게 된다.

본 논문에서는 먼저 PLP 켑스트럼과 음소 인식 시스템에 대해 설명하고 30개 속담 총 106개 단어의 학습에 쓰일 오차역전달법(Error Back Propagation)의 파라메터들(학습율, 모멘텀)을 결정하기 위해 10개 속담만으로 실험을 실시한 결과를 제시하며 이 실험으로부터 얻어진 결과에 따라 선택된 파라메터를 이용하여 30개 속담에 대한 학습을 실시하였다. 학습 효과를 알아보기 위해 인식에 가장 큰 영향을 끼치는 대분류 신경회로망에 대해 학습 데이타의 변화에 따른 학습 결과를 고찰하였다.

Ⅱ. PLP 켑스트럼[3]

음성신호의 세그먼트에 다음과 같은 해밍창(Hamming Window)을 씌운다.

$$W(n)=0.54+0.46\cos\left[\frac{2\pi n}{(N-1)}\right] \tag{1}$$

단, N은 창의 길이이다. FFT(Fast Fourier Transform)를 위하여 25.6ms(256 point)의 시간창을 사용하였으며, 단구간 제곱합을 구하기 위하여 전력 스펙트럼(power spectrum)은 식(2)와 같이 실수성분과 허수성분을 제곱하여 더한다.

$$P(\omega)=\mathrm{Re}[S(\omega)]^2+\mathrm{Im}[S(\omega)]^2 \tag{2}$$

스펙트럼은 다음과 같은 Bark-frequency Ω에 의해 주파수 축을 따라 굴절된다.

$$\Omega(\omega)=6\ln\left\{\frac{\omega}{1200\pi}+\left[\left(\frac{\omega}{1200\pi}\right)^2+1\right]^{0.5}\right\} \tag{3}$$

단, ω는 rad/s의 각속도이다. 이 Bark-Hertz 변환은 Schroeder(1977)[8]에 의해 제안되었다. 굴절된 전력스펙트럼(power spectrum)은 임계대역 마스킹 곡선(critical-band masking curve) $\Psi(\Omega)$와 콘볼루션(Convolution)되며 임계대역곡선(critical-band curve)은 다음과 같이 주어진다.

$$\Psi(\Omega)=\begin{cases}0 & \text{for } \Omega<-1.3\\ 10^{2.5(\Omega+0.5)} & \text{for } -1.3\le\Omega\le-0.5\\ 1 & \text{for } -0.5<\Omega<0.5\\ 10^{-1.0(\Omega-0.5)} & \text{for } 0.5\le\Omega\le2.5\\ 0 & \text{for } \Omega>2.5\end{cases} \tag{4}$$

$\Psi(\Omega)$과 $P(\Omega)$의 이산(discrete) 콘볼루션(convolution)을 하게되면 임계대역 전력스펙트럼 $\Theta(\Omega_i)$이 만들어진다.

$$\Theta(\Omega_i)=\sum_{\Omega=-1.3}^{2.5}P(\Omega-\Omega_i)\Psi(\Omega) \tag{5}$$

표본화된 $\Theta(\Omega)$는 근사된 equal-loudness 곡선에 의해 여과된다.

$$\Xi[\Omega(\omega]=E(\omega)\Theta[\Omega(\omega)] \tag{6}$$

$E(\omega)$는 서로 다른 주파수들에 대하여 사람이 다른 민감도를 갖는것을 근사하고, 약 40dB의 청취 감도를 모방하며 이것에 대한 근사함수는 다음과 같다.

$$E(\omega)=\frac{(\omega^2+56.8\times10^6)\omega^4}{(\omega^2+6.3\times10^6)^2(\omega^2+0.38\times10^9)} \tag{7}$$

식 (7)은 0과 400Hz 사이에서 12dB/oct, 400과 1200Hz 사이에서 9dB/oct; 1200과 3100Hz 사이에서 6dB/oct 그리고 3100Hz 이상에서 0dB/oct의 감쇄율을 가지는 전달함수를 표현한다. 보통의 소리 수준에서는 이러한 근사방법이 5000Hz까지 매우 정확하다. 하지만 더 높은 대역폭을 갖는 응용에 있어서는 5000Hz 이상에서 더 급격한 감쇄율을 갖는 항을 첨가해야한다.

전극점 모형화(all-pole modeling) 전의 처리단계중의 마지막은 3 제곱근 크기 압축(cubic-root amplitude compression)이다.

$$\Phi(\Omega)=\Xi(\Omega)^{0.33} \tag{8}$$

이 처리 단계는 청각의 전력법칙을 근사하고 소리의 세기와 귀가 느끼는 세기의 비선형성을 모방한다. 또한 전극점 모델이 더 적은 차수의 계수를 가지도록 스펙트럼의 변화를 줄이는 작용을 한다.

PLP분석은 LP분석에 비해 계산이 복잡하고 연산량이 많으나 전극점 모델의 결과 스펙트럼은 일반 LP 모델의 스펙트럼에 비해 더 선형적이며[5] 더 낮은 차수의 모델링이 가능하다[5]. 따라서 인공신경망의 입력이 감소하게 되고 학습 속도의 향상에 기여한다.

Ⅲ. 인식시스템

인식 시스템은 10 kHz 샘플링 주파수를 갖는 12 bits A/D, D/A 변환기를 갖는 컴퓨터 및 소프트웨어로 작성된 MLP 분산 신경망으로 구성된다. MLP 신경망으로는 오차역전달 신경망을 사용하였으며 입력, 출력 및 은닉층으로 구성하였다. 신경망 입력으로는 3 ms 마다 256 표본을 취하여 계산된 PLP[3, 4] 켑스트럼(cepstrum)계수[9]와 영교차율, 단구간 에너지 값을 사용하였다.

학습패턴은 인식기의 학습을 위한 목표 값이며 올바른 목표치를 사용하여야 올바르게 학습을 시킬 수 있다[10, 11]. 음성은 연속적인 성질을 가지고 있으며 조음현상의 영향으로 인하여 음의 경계가 분명치 않다. 또한 비슷한 입력에 대해 다른 출력을 요구하는 자체도 인공신경망의 학습시간을 연장시키며 또한 잘못된 입력공간을 구성해 목표치에 수렴하지 못할 가능성이 높다[10, 11]. 그러므로 음소의 경계 부분 즉 천이구간에 해당하는 부분은 학습 목표 구간에서 제외시키고 그 외의 부분을 목표치로 설정하면 학습의 혼돈이 방지될 수 있다. 이와 같이 하지 않으면 음소경계에서는 애매한 출력, 즉 경계 양쪽 음소의 특징을 모두 포함하는 출력을 내도록 학습될 것이므로 전체 학습 및 인식에 상당한 지장을 초래하게 된다. 따라서 자신이 학습에는 쓰이지만 학습 목표로는 설정되지 않는 이런 부분을 20ms 정도 설정하여 학습 목표 설정 시 이를 제외시켰다.

연속음에 나타나는 음소는 전후의 다른 음소와의 조음현상에 의한 변화가 심하다. 현재의 음성 프레임을 인식하기 위해서는 인접한 음소를 같이 참조하여야 자연스럽다. 보통은 이러한 변화를 학습하기 위하여 시간지연신경망을 사용하지만 본 논문에서는 고정 신경망을 사용하고, 입력단의 구성을 시간지연신경망과 다른 형태로 시간적 특성을 수용하였다.

현재 프레임 3ms에 대해 그 전후까지 고려한 특징들의 집합을 구성하는데 그림 1 과 같이 부등간격으로 취해진 각 블럭(block)을 하나의 벡터로 생각하면 한 벡터가 7차 PLP 켑스트럼계수들과 단구간 에너지 및 단구간 영교차율 총 9개의 요소로 구성되었을 경우 7개의 벡터를 사용하면 전체 63개의 데이타가 신경망의 입력으로 구성된다.

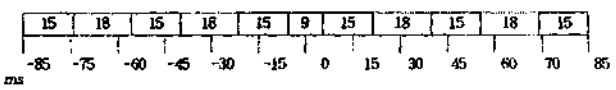

그림. 1. 입력벡터의 요소 구성
Fig. 1. The distribution of PLP features

네모칸 안의 숫자는 각 블럭의 시간폭을 나타낸다. 실선은 PLP가 평균되어지는 구간의 길이를 나타내며, 점선은 사용되지 않는 구간의 길이이다. 9는 현재의 프레임(3ms) 및 전 프레임(3ms), 후 프레임(3ms), 총 3개의 프레임으로 한 블럭을 형성한 것을 나타내며 세 프레임에서 구한 특징들이 평균되어 한 특징 벡터로 나타내어진다. 15는 5개의 프레임(총15ms)으로 한 블럭을 형성한 것을 나타내며 다섯 프레임에서 구한 특징들이 평균되어 한 특징 벡터로 나타내어진다. 18은 사용되지 않는 구간인데 6개의 프레임 즉 18ms 길이이다.

분산 신경망은 그림2와 같이 입력된 특징값에 따라 8개의 군으로 분류해주는 음소 대분류 신경회로망과 각 군에서 음소를 인식하는 음소 인식 신경회로망으로 구성되어 있다. 속담에 쓰인 음소는 총 34개로서 유성 자음인 ㄱㄷㅂ군, 무성자음인 ㅋㅌㅍ캉군, 비음인 ㄴㄹㅁㅇ군, 된소리인 ㄲㄸㅃㅆㅉ군, 마찰음 및 파찰음인 ㅅㅈㅊ군,

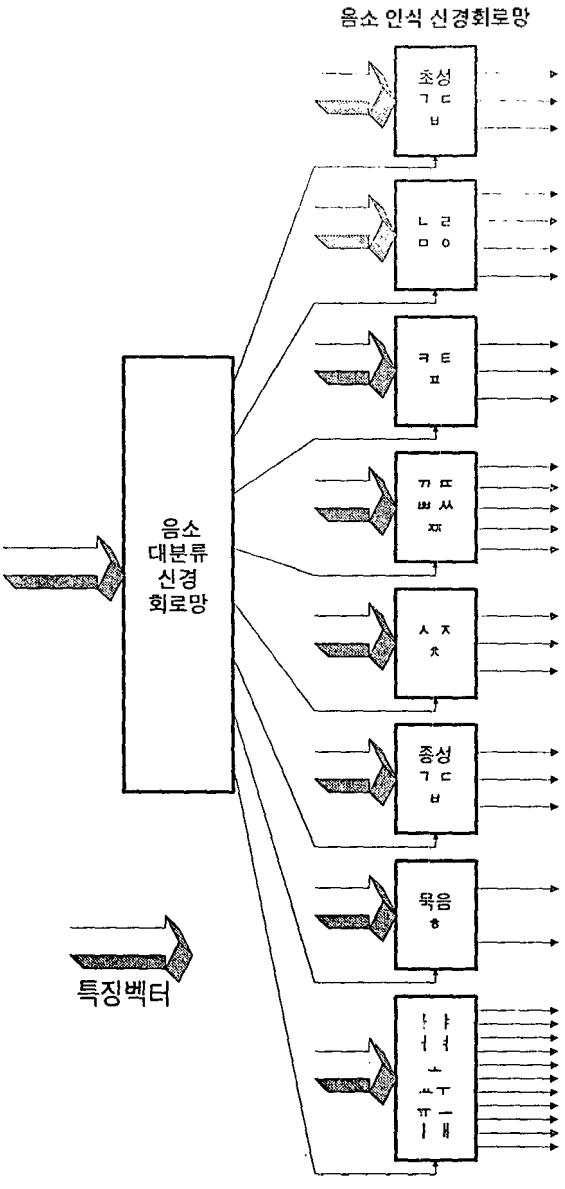

그림. 2. 분산 신경망의 구조
Fig. 2. The structure of distributed neural network

종성ㄱㄷㅂ군 묵음 및 하군, 모음군으로 구별하였다. 각 신경망의 입력층의 뉴런의 갯수는 63개이며 은닉층의 뉴런의 갯수는 20개를 사용하였으며 여기에 바이어스를 추가하였다. 출력층의 뉴런의 갯수는 음소 대분류 신경회로망은 8개이며 각 음소 인식 신경회로망은 인식하고자 하는 음소의 갯수에 따라 결정된다.

사용된 자모는 묵음까지 34개이다.

Ⅳ. 실험 및 검토

한국 속담 30개를 나열하면 다음과 같다.(괄호안은 발음대로 적은 것)

1. 가자니 태산이요 돌아서자니 숭산이라.(가자니 태사니요 도라서자니 숭사니라)
2. 고래 싸움에 새우등 터진다.(고래 싸우메 새우등 터진다)
3. 구르는 돌에는 이끼가 끼지 않는다.(구르는 도레는 이끼가 끼지 안는다)
4. 낙숫물이 댓돌을 뚫는다.(낙쑨무리 대또를 뚤는다)
5. 눈뜨고 도둑 맞는다.(눈뜨고 도둑 만는다)
6. 달면 삼키고 쓰면 뱉는다.(달면 삼키고 쓰면 밴는다)
7. 뚝배기보다 장맛이다.(뚝빼기보다 장마시다)
8. 말똥도 세번 굴러야 제자리에 선다.(말똥도 세번 굴러야 제자리에 선다)
9. 목욕한 사람은 옷도 털어입는다.(모교칸 사라믄 온또 터러임는다)
10. 미련은 먼저나고 슬기는 나중난다.(미려는 먼저나고 슬기는 나중난다)
11. 백지장도 맞들면 낫다.(백찌짱도 마뜰면 낟따)
12. 부자는 망해도 삼년 먹을것이 있다.(부자는 망해도 삼년 머글거시 인따)
13. 사향노루는 배꼽 때문에 죽는다.(사향노루는 배꼽 때무네 중는다)
14. 새벽달 보자고 초저녁부터 기다린다.(새벽딸 보자고 초저녁부터 기다린다)
15. 숲이 깊어야 도깨비가 모인다.(수피 기퍼야 도깨비가 모인다)
16. 여우를 피하면 호랑이를 만난다.(여우를 피하면 호랑이를 만난다)
17. 옆구리 찔러 절받기(엽꾸리 찔러 절받끼)
18. 육갑도 모르고 산통 깬다.(육깝또 모르고 산통 깬다)
19. 지성이면 감천이다.(지성이면 감처니다)
20. 짚신장수 헌신만 신는다.(집씬장수 헌신만 신는다)
21. 참새가 방앗간 그냥 지나가랴.(참새가 방알깐 그냥 지나가랴)
22. 천리길도 한걸음부터(철리낄도 한거름부터)
23. 칼로베고 소금친다.(칼로베고 소금친다)
24. 큰방죽도 개미구멍으로 무너진다.(큰방죽도 개미구멍으로 무너진다)
25. 태산을 넘으면 평지를 본다.(태사늘 너므면 평지를 본다)
26. 토끼를 잡으면 사냥개를 죽인다.(토끼를 자브면 사냥깨를 주긴다)
27. 포도청 문고리도 빼겠다.(포도청 문꼬리도 빼겠다)
28. 핑계없는 무덤없다.(핑게엄는 무덤업따)
29. 흐르는 물도 떠주면 공덕이다.(흐르는 물도 떠주면 공더기다)
30. 호미로 막을걸 가래로 막는다.(호미로 마글껄 가래로 망는다)

속담 30개는 문법적인 띄어쓰기와 무관하게 발음상 우리나라 단어의 운율인 3,4 구조가 나오도록 띄어쓰기가 되었고 띄어쓰기 된 부분에서 인위적으로 잠깐 멈춘 후 발음하도록 하였다. 띄어쓰기 된 부분을 기준으로 한 단어씩 구분할 때 총 106 단어가 된다. 20대 남성 한 사람의 음성을 실험실에서 디지탈 오디오 테이프 녹음기로 녹음하여 컴퓨터에 입력하였다. 학습은 분산 신경망을 사용하여 음소 대분류 신경회로망을 학습시킨 후 각각의 군에 따라 음소 인식 신경망을 학습시켰다.

1. 학습 변수

원활한 학습을 위해서는 학습 변수의 선택이 중요하다. 즉 학습율과 모멘텀의 선택이 그것인데 학습율 0.5와 모멘텀 0.5를 기준으로 각각을 ±0.2 씩 변화시켜 학습율 0.5일 때 모멘텀 0.7, 0.5, 0.3일 경우와 모멘텀 0.5일 때 학습율 0.7, 0.3인 경우의 5가지 경우에 대해서 학습율과 모멘텀이 어느 정도일 때 가장 빨리 수렴하는지를 평균 오차 0.001, 최대 학습 회수 2000회를 기준으로 실시하였다. 학습에 쓰이는 음성은 20대 남성 1명의 음성이며 학습 시간이 많이 걸리므로 속담 1 번부터 10 번까지에 대해 실험을 실시하였다. 실험에 쓰인 기계는 LINUX OS를 쓰는 75㎒ 펜티엄 PC와 CRAY 수퍼 컴퓨터,두 종류이

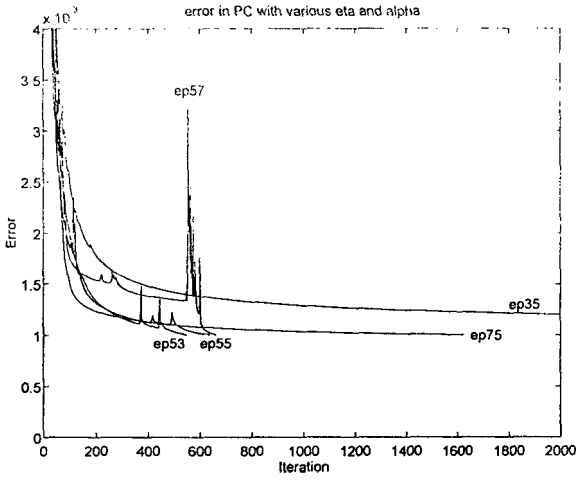

그림. 3. 여러가지 학습율과 모멘텀으로 PC에서 구한 10개 속담 학습 평균 오차도

Fig. 3. Mean errors with various learning rate and momentum in PC

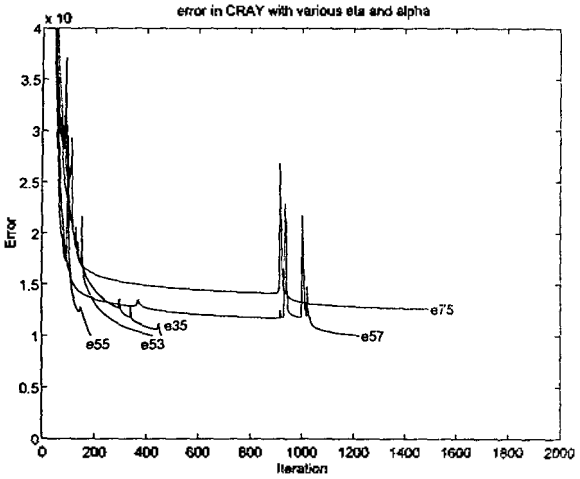

그림. 4. 여러가지 학습율과 모멘텀으로 CRAY에서 구한 10개 속담 학습 평균 오차도

Fig. 4. Mean errors with various learning rate and momentum in CRAY

다. PC와 컴퓨터는 속도 차이 뿐만 아니라 부동소숫점 계산 자리수가 4바이트와 8바이트로 각각 차이가 나므로 수많은 반복 계산을 하는 신경회로망에서는 그 차이가 상당하다. 그림 3에 PC 에서 수행한 결과가, 그림 4에 수퍼 컴퓨터에서 수행한 결과가 나타나 있다.

PC에서는 학습율 0.5와 모멘텀 0.3이 가장 빨리 수렴하지만 CRAY에서는 학습율 0.5와 모멘텀 0.5일 때 가장 빨리 수렴하였다. PC에서는 학습율이 0.3 모멘텀이 0.7일 경우에 2000회까지 가더라도 수렴하지 않는 것으로 나타났으나 CRAY에서는 학습율 0.7 모멘텀 0.5일 경우로 나타났다. 두 경우 다 모멘텀이 클 경우(0.7)에 학습도중에 오차값이 크게 변화한 후에 변화 전보다 오차가 현저히 줄어드는 현상이 나타났다. 대개 학습율과 모멘텀이 클수록 반복 초기에 오차가 빨리 줄어드는 경향이 있으며 그림 3과 그림 4에서도 이런 현상을 볼 수 있다. 어떤 값이 가장 적절한지는 정확히 결론내릴 수 없고 데이타의 갯수가 바뀌면 수렴 속도 및 횟수도 다 다르게 나타날 것이나 학습율 및 모멘텀이 너무 크거나 작을 경우보다는 그 중간값이 학습율 0.5 모멘텀 0.5가 CRAY에서 가장 빨리 수렴하고 PC에서는 두번째로 빨리 수렴하는 결과를 보여주었고 또한 학습에 쓰이는 데이타가 바뀌더라도 이 값이 적절할 것이라고 예상된다. 따라서 이후의 실험에서는 학습율 0.5와 모멘텀 0.5를 사용하였다.

2. 학습 데이타와 인식율

30개 속담 인식 실험은 화자 종속 인식 방식으로 20대 남자 한명의 음성을 각 속담에 대해서 다섯번씩 받아 이중에 네집단(30개 속담 음성 한번 받은 것을 한집단이라 하자)을 학습에 사용하고 한집단은 학습에서 제외시켜 학습에 참여하지 않은 데이타의 인식에 사용하였다. 사용 집단수의 증가와 이에 따른 학습효과와 인식 효과를 알아 보기 위해 먼저 한집단(30개 속담)을 학습시킨후 학습에 참여한 데이타의 음소 인식율과 학습에 참여하지 않은 데이타의 음소 인식율을 알아보고 두번째는 두집단(60개 속담)에 대해서 세번째는 세집단(90개 속담) 네번째는 네집단(120개 속담)에 대해서 실험하였다. 특징값으로는 7차 PLP와 영교차율, 에너지를 사용하고 입력벡터 구성과 학습 목표치등 모든 것이 제안된 방법대로 실시되었다. 학습율 0.5 모멘텀 0.5를 사용하고 최대 반복 횟수를 2000으로 제한하였다. 평균 오차를 0.001로 주었지만 대개 이 값에 수렴하지 못하고 2000회에 걸려서 학습을 종료하였다. 프레임 인식률은 그 후에 진행될 음소 인식, 단어 인식, 문장 인식등에 큰 영향을 끼친다. 분산 신경회로망 내에서는 음소 대분류 신경회로망의 정확한 음소 분류가 전체 인식율에 결정적인 영향을 주므로 음소 대분류 신경회로망을 대상으로 실험을 실시하였다. 먼저, 학습에 참여한 데이타에 대한 인식율과 학습에 참여하지 않은 데이타에 대한 인식율 변화 추이를 그림 5를 통해 살펴본다. 학습을 완료한 후 학습에 의해 결정된 가중치로 학습에 참여한 음성 데이타를 대상으로 인식을 실시한 결과 학습 데이타가 적을 때에는 학습이 거의 완벽하게 되었으나 데이타의 수가 늘어날수록 학습의 부담이 증가하고 이에 따라 인식율이 조금씩 감소하였다. 이에 반해 학습에 참여하지 않은 데이타에 대한 인식에서는 오히려 인식율이 현저히 증가하는 현상을 보여 주었다. 이것은 개인의 음성간에도 여러가지 요인으로 인해 변화가 심하다는 것을 단적으로 보여주고 있으며 가능하면 여러가지 경우를 다 학습시키는 것이 인식율 향상에 도움을 준다는 것을 나타내고 있다. 그러나 그 증가 양상이 지수 함수적 형태를 띠고 있어서 한 음성만 학습에 사용한 경우보다 두 음성을 사용한 경우가 현저한 인식율 증가를 보여주지만 그 후 부터는 완만한 증가를 나타낸다. 따라서 무작정 많은 데이타로 학습 시키는 것보다는 어느 정도 타협선이 필요하다.

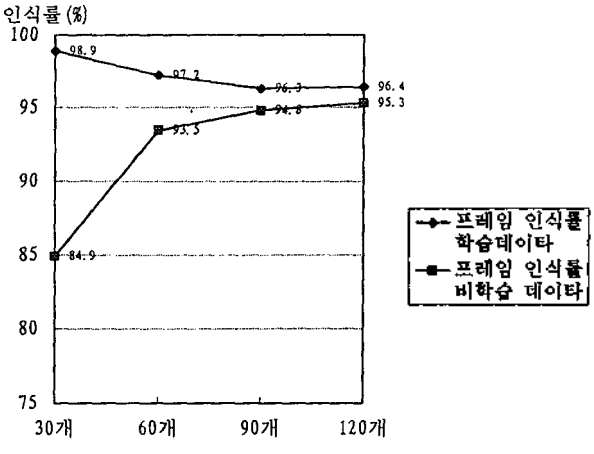

그림. 5. 학습 참여 데이타의 증가에 따른 인식률 추이

Fig. 5. Trends for the recognition rate as the data increases

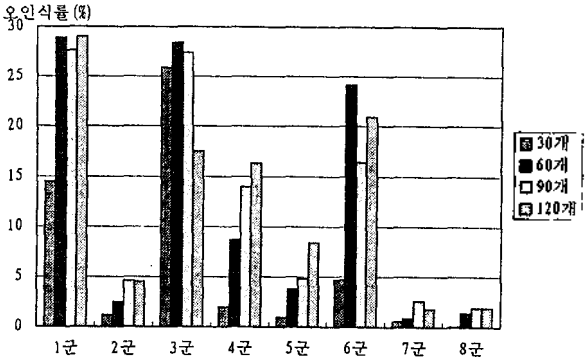

그림. 6. 학습에 사용된 데이타에 대한 학습 데이타 증가에 따른 각 군별 오인식율 그래프

Fig. 6. Graph for the ratio of erroneous recognition in each group for the data being used in the learning as the learning data increases

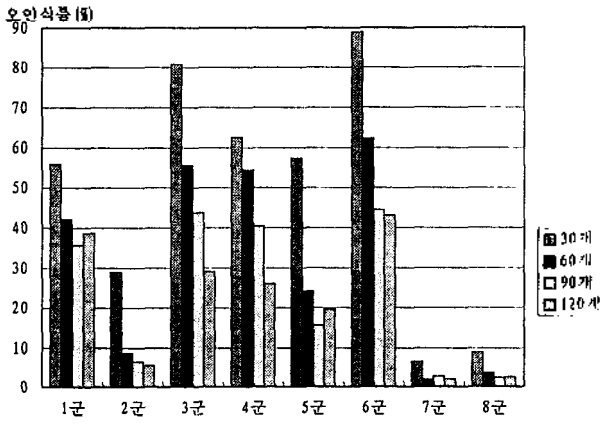

그림. 7. 학습에 사용되지 않은 데이타에 대한 학습 데이타 증가에 따른 각 군별 오인식율 그래프

Fig. 7. Graph for the ratio of erroneous recognition in each group for the data being unused in the learning as the learning data increases

앞서는 전체적인 인식율을 기준으로 결과를 살펴보았는데 이제 각 군별 인식율을 기준으로 학습 데이타의 증가에 따라 각 군의 프레임 오인식율은 어떻게 변화하는지 살펴본다. 그림 6에는 학습에 참여한 데이타에 대한 각 군별 오인식율이 막대 그래프로 나타나 있고 그림 7에는 학습에 참여하지 않은 데이타에 대한 각 군별 오인식율이 막대 그래프로 나타나 있다.

초성 ㄱㄷㅂ를 1군 ㄴㄹㅁㅇ을 2군 ㅋㅌㅍ를 3군 ㄲㄸㅃㅆㅉ를 4군 ㅅㅈㅊ를 5군 종성 ㄱㄷㅂ를 6군 묵음 및 ㅎ을 7군 모음 ㅏㅑㅓㅕㅗㅛㅜㅠㅡㅣㅐ를 8군으로 정하였다. 두 경우 다 자음의 오인식율이 상당히 높고 모음 및 유성 자음 그리고 묵음 및 ㅎ군인 7군이 오인식율이 낮다. 자음 인식의 어려움은 이미 잘 알려져 있는 바이고 이것이 이번 실험에서도 여실히 증명되었다.

학습에 참여한 데이타의 인식에서는 대체로 데이타의 갯수가 증가할수록 오인식율이 증가하는데 전체 인식율의 저하에 자음의 오인식이 크게 기여하고 있음을 알 수 있다. 학습에 참여하지 않은 데이타에 대한 오인식율은 학습 데이타가 증가할 수록 감소되는 것을 알 수 있다. 자음의 오인식이 높은 것처럼 학습 데이타의 증가에 따른 오인식의 개선도 자음에서 현저하게 나타났다.

어떤 군이 어느 군으로 오인식 되었는지 알기 위해 네 집단으로 학습된 신경망으로 학습에 참여하지 않은 데이타에 대한 인식을 실시하고 이에 대한 혼돈표(confusion matrix)를 표 1에 나타내었다.

자음은 유성 자음인 ㄴㄹㅁㅇ을 빼 놓고는 그 길이가 상당히 짧으므로 전체 문장의 프레임 갯수에서 자음이 차지하는 갯수가 상대적으로 적다. 따라서 이들의 학습 기회도 적어서 학습 기회 불균형이 생기게 되며 자음군의 오인식도 증가하게 된다. 각 군의 오인식율은 자음에서 현저히 크게 나타나는 것을 알 수 있는데 초성 ㄱㄷㅂ군과 종성 ㄱㄷㅂ군이 각각 2위, 1위를 차지하고 있다. 유성음인 2군과 8군이 적게 나타나며 대부분 묵음으로

표 1. 네 집단으로 학습된 대분류 신경회로망의 혼돈표

Table 1. Confusion matrix for the large grouping neural network learned from four sets of data

오인식군 / 소속군	1	2	3	4	5	6	7	8	불인식	소계	인식대상 프레임수	오인식율 (%)
1	0	34	58	3	30	0	3	23	36	187	485	38.6
2	4	0	0	0	2	2	45	87	21	161	3,016	5.3
3	2	3	0	5	8	0	0	8	18	44	151	29.1
4	0	0	5	0	29	0	8	1	16	59	226	26.1
5	12	8	5	52	0	0	12	21	35	145	738	19.6
6	0	12	0	0	0	0	30	5	11	58	135	43.0
7	12	7	5	11	11	22	0	33	51	152	7,845	1.9
8	2	104	2	2	1	0	54	0	43	208	8,762	2.4
(인식된 프레임 수 : 20344/(전체 인식대상 프레임 수 : 21,358) = 95.3%												

구성된 7군 역시 안정된 데이타로 학습되므로 오인식율이 낮음을 알 수 있다.

각 군의 오인식 분포 비율을 나타내는 그림 8을 보면 유성 파열음인 1군은 31%가 같은 무성 파열음인 3군으로 오인식되었다. 이것은 같은 파열음 계열로서 음성의 특징이 비슷하다고 볼 수 있는데 거꾸로 3군에서는 이런 현상이 관찰되지 않는다. 또한 2군(18%) 및 8군(12%)으로의 오인식도 많은데 이것은 ㄱㄷㅂ중에서 유성음인 것과 무성음인 것을 구별하지 않은 결과 유성음 ㄱㄷㅂ가 유성음군인 2군 및 8군으로 오인식된것으로 추정할 수 있다. 5군으로의 오인식(16%)도 꽤 존재하는 것은 역시 마찰음 및 파찰음 계열과 크게 구별되지 않는 자음의 공통적 특징인 잡음성 때문일 것이다. (2군은 55%가 8군으로 오인식되는데 유성음의 특징 때문으로 판단할 수

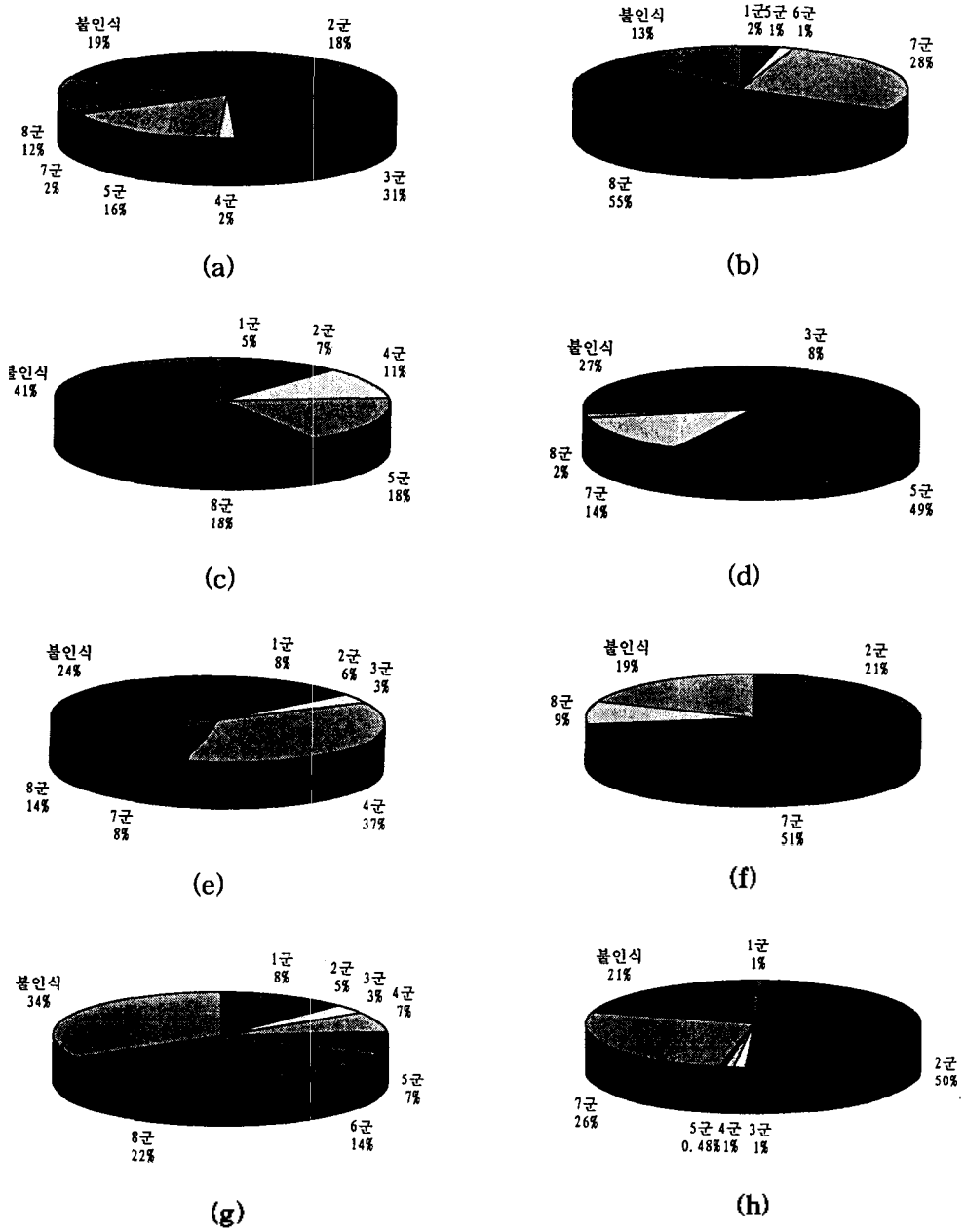

그림. 8. 각 군의 오인식 분포율 도표
(a) 1군 (b) 2군 (c) 3군 (d) 4군
(e) 5군 (f) 6군 (g) 7군 (h) 8군

Fig. 8. The distribution ratio chart for the erroneous recognition in each group
(a) 1 group (b) 2 group (c) group (d) group
(e) 5 group (f) 6 group (g) 7 group (h) 8 group

있다. 3군은 짧은 시간만 나타나므로 41%가 인식되지 않았다. 4군에서는 49%가 5군으로 5군에서는 37%가 4군으로 오인식되는 것을 보면 ㅆ와 ㅉ가 5군의 ㅅ 및 ㅊ와 같은 마찰음 및 파찰음 계열인 것과 무관하지 않다. 종성 ㄱㄷㅂ의 7군은 묵음 및 ㅎ으로 분류되는 8군으로 가장 많이 오인식되어서 51%를 나타내고 있는데 종성 ㄱㄷㅂ은 에너지가 급격히 감소되는 부분이므로 에너지가 거의 없는 묵음으로 혼돈될 가능성을 애초에 지니고 있다. 8군은 2군으로 가장 많이 오인식 되어서 50%를 나타내고 있는데 2군이 8군으로 가장 많이 오인식 된 것과 무관하지 않아서 유성음이라는 공통점에 기인한 것으로 해석할 수 있다.

Ⅵ. 결 론

한국어 속담 30개에 대해 20대 남자가 5번 발음한 음성 데이타에 대해 분산 신경망을 이용하여 프레임별 음소 인식을 실시하였다. 대규모의 데이타를 학습시키기 위해 일반적으로 사용하는 오차역전파 인공 신경망을 기능을 분산시켜 음성 특징이 비슷한 군으로 분류하는 신경망과 분류된 음소를 인식하는 신경망으로 구분하여 학습 및 인식을 실시하여 학습의 부담을 줄였다. 학습에 쓰이는 파라메터를 결정하기 위해 10개 속담을 사용하여 실험을 실시하여 학습율과 모멘텀을 결정하고 결정된 학습율과 모멘텀을 사용하여 30개 속담에 대한 인식을 실시하였다. 음소 인식에서 가장 중요한 열쇠를 쥐고 있는 음소 대분류 신경회로망의 음소 인식율이 학습에 쓰이는 데이타의 갯수에 따라 어떻게 변화하는지를 살펴보았다. 학습 데이타가 증가할수록 학습에 참여한 데이타의 인식율은 떨어졌지만 학습에 참여하지 않은 데이타의 인식율이 올라감을 실험을 통하여 확인하였다. 학습에 참여하는 데이타의 숫자가 늘어날수록 비학습 데이타에 대한 인식율은 84.9%에서 95.3%까지 올라갈 수 있었다.

음성 인식의 용도는 현재 제한된 어휘에 대해 한정된다. 본 연구는 단어가 중복되지 않는 30개 속담, 106개 단어를 문장 단위로 단어 사이에 묵음을 두어 발음하여서 이를 프레임별 음소 인식하였으며 제안된 분산 신경망의 대분류 신경망의 기능과 역할에 촛점을 맞추어 실험하였다. 앞으로 인식된 음소로 단어를 인식하고 단어로부터 문장을 인식하게 되면 간단한 문장에 의한 음성 명령 및 인식기의 구현이 가능하리라 기대된다.

참 고 문 헌

1. 김형순, "Keyword Spotting 기술," 한국통신학회지, vol. 11, no. 9, pp. 57-66, pp. 57-66, 1994. 9.
2. Richard D. Peacocke and Daryl H. Graf, "An Introduction to Speech and Speaker Recognition," COMPUTER, vol. 23, no. 8, pp. 26-23, 1990. 8.
3. H. Hamansky, "Perceptual Linear Predictive(PLP) analysis of speech" J. Acoust. Soc. Am., 87(4) : 1738~1752, April 1990
4. Rik D. T. Janssen, Mark Fanty and Ronald, "Speaker Independent Phonetic Classification in Continuous English Letters," INNS vol. 2, pp. 801 ~808, 1991
5. H. Harmanskey, Kazuhiro Tsuga, Shozo Makino, and Wakita., "Perceptually Based Processing In Automatic Speech Recognition," ICASSP, pp. 1971-1162, 1986
6. L. R. Rabiner, S. E. Levinson and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent, Isolated Word Recognition," The Bell System technical Journal, Vol. 62, No. 4, April 1983
7. L. R. Rabiner, J. G. Wilpon and F. K. Soong, "High Performance Connected Digit Recognition Using Hidden Markov Models," IEEE Trans. Acoust., Speech, Signal Processing, Vol. 37, No. 8, Aug. 1989
8. M. R. Schroeder, "Recognition of Comples Acoustic Signals," Life Science Research Report 5, edited by T. H. Bullock (Abakon Verlag, Berlin), p. 324
9. S. Saito and K. Nakata, *Fundamentals of Speech Signal Processing*, 1985, Academic Press
10. T. W. Parsons, *Voice and Speech Processing*, pp. 59-81, 1986, McGraw Hill Inc.
11. P. D. Wasserman, *Neural Computing:Theory and Practice*, 1993, Van Nostrand Reinhold New York

▲김 선 일

1960年 3月 19日生
1983年 2月 : 아주대학교 전자공학과 졸업(공학사)
1985年 2月 : 아주대학교 대학원 전자공학과 졸업(공학석사)
1994年 2月 : 아주대학교 박사과정 수료
1985年 3月 ~1990年 2月 : 한국기계연구소 자동제어실 연구원
1990年 3月 ~1990年 8月 : 한국기계연구소 자동제어실 선임연구원
1990年 8月 ~現在 : 거제전문대 전자과 교수

▲이 행 세
1943年 8月 29日生
1966年 2月 : 전북대학교 전기공학과 졸업(공학사)
1972年 2月 : 서울대학교 대학원 전자공학과 졸업(공학석사)
1984년 : 고려대학교 대학원 전자공학과 졸업(공학박사)
1973年 2월 ~現在 : 아주대학교 전자공학과 교수