오차 역전파 알고리즘을 갖는 MLP를 이용한 한국 지명 인식에 대한 연구

A Study on the Spoken Korean Citynames Using Multi-Layered Perceptron of Back-Propagation Algorithm

송 도 선*, 이 재 건**, 김 석 동***, 이 행 세****

(Do Sun Song, Jaegheon Lee, Seok Dong Kim, and Haing Sei Lee)

요 약

이 논문은 오차역전달(error back-propagation) 알고리듬을 갖는 다층구조 퍼셉트런(Multi-Layered Perceptron)을 사용하여 우리말 단어음성을 화자종속으로 기계 인식하는 실험에 관한 연구 결과다.

대상단어는 시외 자동전화 지역번호표에서 임의로 선택한 50개 지역명이며, 이 중 43개는 2음절로 구성되어있고 나머지 7개는 3음절이다. 단어를 음소나 음절별로 분리(segmentation)하지 않고, 단어의 각 부분에서 골고루 추출된 특징성분을 신경망에 입력하는 방법을 사용했다. 그렇게 함으로써 발음지속시간에 관계없는 결과를 얻을 수 있으며, 이 때 사용된 특징 성분은 선형예측분석으로 구해진 PARCOR계수다. 전체학습과 구분학습의 비교, 프레임 갯수와 PARCOR차수에 대한 인식률의 의존도, 중간층 뉴런의 갯수에 대한 인식률의 변동, 그리고 출력층 뉴런의 구성 방법에 따른 비교 등 4가지 실험을 통하여 가장 최량의 조건을 찾아보고자 하였다.

이 연구를 발전시킨다면 실시간의 화자독립 소규모어휘 음성인식이 가능해질 것으로 보인다.

ABSTRACT

This paper is about an experiment of speaker-independent automatic Korean spoken words recognition using Multi-Layered Perceptron and Error Back-propagation algorithm.

The object words are 50 citynames of D.D.D local numbers. 43 of those are 2 syllables and the rest 7 are 3 syllables. The words were not segmented into syllables or phonemes, and some feature components extracted from the words in equal gap were applied to the neural network. That led independent result on the speech duration, and the PARCOR coefficients calculated from the frames using linear predictive analysis were employed as feature

*중경공업전문대학교 전자과 교수
**전자통신 연구소 이동통신 기술 연구단
***호서대학교 전자계산학과 교수
****아주대학교 전자공학과 교수
접수일자 : 1994년 1월 30일

components. This paper tried to find out the optimum conditions through 4 differerent experiments which are comparison between total and pre-classified training, dependency of recognition rate on the number of frames and PAROCR order, recognition change due to the number of neurons in the hidden layer, and the comparison of the output pattern composition method of output neurons. As a result, the recognition rate of 89.6% is obtaimed through the research.

I. 서 론

인간사이에서 가장 중요한 통신방법은 바로 언어다. 언어에는 구어(口語)와 문어(文語)가 있으며, 이는 각각 음성에 의한 청각적 기호정보 전달수단과 문자에 의한 시각적 기호정보 전달수단에 해당된다.

음성언어는 생성되자마자 바로 소멸되어버리며, 보조장비를 사용하지 않는한 그 전파범위가 상당히 제한되어있는 반면, 인간의 사고속도에 비해 생성되는데에 걸리는 시간이 그다지 많지 않으므로 즉각적인 통신방법으로서 많이 사용된다.

최근 각종 정보처리기술의 발달과 함께 인간과 기계사이의 통신수단으로서 음성언어를 이용하는 방안이 구체적으로 연구되고 있다[1-6]. 실로 지금까지 데이타형 정보를 처리하는 기술은 급격하게 발달되어온 데에 반해, 그 정보의 입출력 형태는 문자, 화상, 음향 등으로 극히 제안되어 있었던 것이 사실이며, 그중에서도 인간이 가진 정보를 기계에 입력하는 방법은 펀치카드와 키보드가 고작이었다. 따라서 어떤 문제를 처리하는데에 있어서 실제로 문제를 해결하는데에 걸리는 시간과 비용보다 문제를 기계에 입력시키는데에 걸리는 그것이 더 큰 경우가 자주 발생하며, 또 어느정도 숙련된 사람이 아니면 장비를 사용하는데에 어려움을 가질 수밖에 없었던 것이다. 이런 경우에 기계가 음성언어를 알아듣도록 하는 음성인식기술은 대단히 중요한 연구 과제가 아닐 수 없다[7].

한편, 현대과학기술의 가장 큰 목표중의 하나는 인간의 지능과 맞먹는, 혹은 더 나아가서 인간보다 더 높은 지능을 갖는 기계를 발명하여 인간을 돕게 하는 것이다. 지능은 한마디로 요약할 수 없는 복합적인 기능이지만 사람이 쓰는 말뜻과 글뜻을 이해하는 기능도 지능의 중요한 요소인 것이다.

인간이 말을 알아듣는 것과 비슷한 정도로 우수한 성능을 가진 음성인식기술이 실현되면 문서작성시에 드는 수고를 훨씬 덜 수 있게 될 뿐만 아니라 자동 국제언어 번역기, 장애자 보조장치, 자동응답, 전화시스템, 공장자동화 등 그 응용범위가 무한히 넓다. 또한 고밀도의 정보 송수신시대에 있어서 정보압축기술도 상당히 중요한 연구과제임에 틀림없는데, 사실 음성신호의 압축방법중에서 가장 고효율의 압축법은 바로 음성을 인식해서 데이타형 정보로 전송하는 것이다[8].

음성인식에 대한 연구는 음성의 전송 및 합성기술과 더불어 같은 역사적 배경을 가지고 발전되어왔으며, 디지틀 신호처리기술의 발달에 힘입은 바 크다[9].

1952년 벨연구소(Bell Lab)에서 처음으로 포먼트(formant) 주파수를 이용하여 숫자음을 인식하는데에 부분적으로 성공한 이후로 간간히 연구결과가 소개되다가, 1970년대 초반에 미국에서 행해진 DARPA 프로젝트에 음성이해에 대한 연구가 포함되면서 본격적인 연구가 시작되었다. 70년대 중반에는 일반적인 문제해결 알고리듬(strategy)에 대한 상대적으로 특수한 지식(knowledge)의 중요성이 알려지게 되면서 음성 그 자체와 청각기관의 구조 및 기능에 대한 연구로부터 시작되는 상향식 접근방법(bottom-up approach)을 택하는 계기가 마련된다.

벨연구소의 Rabiner, IBM의 Jelinek, CMU의 Waibel, MIT의 Zue, Klatt, Lippmann 등의 연구자들이 DTW(Dynamic Time Warping), HMM(Hidden Markov Model), VQ(Vector Quantization) 등의 방법을 이용하여 음성인식 연구를[12, 13, 14] 행한 바 있고, 1984년 이후로 Kohonen, Lippmann, Anderson, Lang, Waibel, Sawai, Miyatake 등이 ANN(Artificial Neural Network)을 [10, 11, 15] 이용한 음성인식에서 상당한 성과를 얻고있으며, 현재 퍼지이론(fuzzy theory)을 도입한 연구[22], HMM의 비테르비(Viterbi) 알고리듬을 신경망에 결합한 복합기술연구 등이 활발히 진행되고 있다[20, 21].

음성정보를 인식하는 알고리듬을 개발하는데에 몇가지 어려운 점이 아직도 해결되지 않고 있는 실정이다[16]. 의학과 생물학의 발전으로 청각기관의 구조와 기능은 이미 많은 부분이 알려져 있지만, 그 청각기관에서 내보내는 신호를 해석하는 두뇌의 기능은 여전히 신비에 싸여있다[17, 18, 19].

같은 음운이라도 사람에 따라 발성된 음성신호는 제각기 다르며, 같은 사람이 발음한 동일음운일 때 조차도 발음때마다 항상 다르다. 심리상태에 따른 음성 변화도 상당한 것이다.

주변의 소음으로부터 의미있는 음운만을 골라내는 것 또한 상당히 세밀한 기술을 요하며, 여러 음운이 연속하여 발음될 때 발생하는 구개음화, 종성규칙, 자음동화, 자음축약, 연음법칙, 경음화, 음운첨가, 예외발음 등의 조음현상(coarticulation)이 정확한 음운의 식별을 방해하는 일도 있어서 음절간 혹은 단어간의 경계 분리(segmentation)가 쉽지 않다. 현재 한 음절에서 음소간의 경계를 분리하는 문제도 완전히 해결되지 않고있는 상태다.

또, 음성 인식이 가능해지더라도 인간이 하는 것처럼 음성의 내용을 이해하기 위해서는 음성학(phonetics), 음운학(phonology), 음운배열론(phonotactics), 시형론(prosodic), 구문론(syntax), 의미론(semantics), 어형론(morphology) 등에 대한 지식을 갖고있지 않으면 안된다.

이러한 점들이 실용적인 기계의 음성인식을 가로막고 있는 요소들로 작용한다.

Ⅱ. 실험 내용 및 실험 방법

본 연구는 한국어 단어음성을 그 실험 대상으로 하고 있으며, 그 대상으로 한국의 지명 몇가지를 선택했다. 시외자동전화(D.D.D) 지역번호표의 지역명은 대부분 2음절로 이루어져 있기때문에, 기초적 단어 인식 실험을 위한 대상으로 아주 적절하다. 하지만 '-주(州)' 또는 '-산(山)'과같이 한국의 많은 지명에 공통적으로 들어가는 음절이 몇가지 있기때문에 단어들의 패턴 특징 거리(pattern feature distance)가 가깝다는 것이 인식 실험에 있어서 부적당한 요인이 될 수도 있다. 임의로 50개 지역명을 선택하였으며, 선택된 지명중에서 43개는 2음절 단어이고, 나머지 7개는 3음절이다. [표 1]에 실험에 선택된 50개 지명이 실려있다.

비교적 주변 소음으로부터 차단된 환경에서 위의 지역명이 1명의 20대 남자의 음성으로 한 지역명당 10번씩 발음되어 DAT(Digital Audio Tape)에 녹음되었고, 이것이 아날로그-디지탈 변환(Analog-to-Digital Conversion)되어 PC의 하드디스크(hard disk)에 저장되었다. ADC과정은 샘플링주파수(sampling rate) 10 kHz, 양자화해상도(quantization level)12bits/sample의 정밀도로 DT2801 ADDA(Analog-to-Digital/Digital-to-Analog Converter)를 통하여 행해졌다.

표 1. 실험에 사용된 50개 지역명
Table 1. 50 citynames used for the experiment

1. 서 울	11. 문 산	21. 여 주	31. 설 악	41. 태 백
2. 부 산	12. 미 금	22. 연 천	32. 양 구	42. 평 창
3. 대 구	13. 발 안	23. 오 산	33. 원 당	43. 홍 천
4. 광 주	14. 수 원	24. 용 인	34. 보 은	44. 남양주
5. 대 전	15. 시 흥	25. 의 왕	35. 영 월	45. 의정부
6. 강 화	16. 성 남	26. 일 산	36. 원 주	46. 장호원
7. 고 양	17. 안 산	27. 파 주	37. 인 제	47. 주문진
8. 과 천	18. 안 성	28. 하 남	38. 정 선	48. 연무대
9. 광 명	19. 안 양	29. 화 성	39. 철 원	49. 장승포
10. 화 천	20. 양 주	30. 도 계	40. 춘 천	50. 동광양

각 파일에 저장된 데이타(raw data)에는 단어음성의 앞과 뒤에 음성이 들어있지 않은 무음구간(silence interval)이 들어있으므로 이 부분을 제거하고 음성구간만을 뽑아내기위해 음성구간 검출 알고리듬이 필요하다. 본 실험에 사용된 음성구간 검출 알고리듬은 [그림 1]에서 보는 바와 같다.

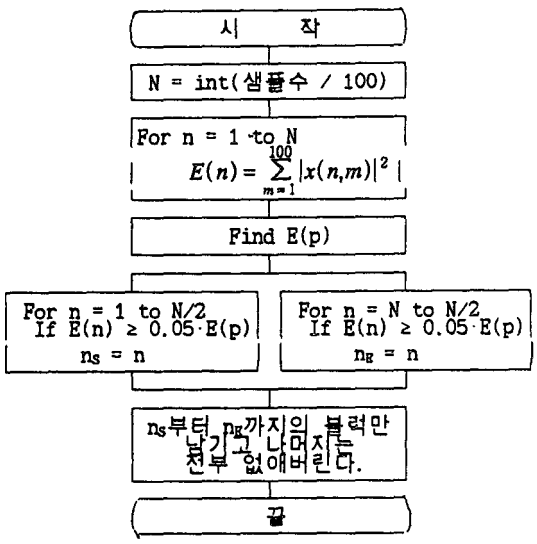

그림 4.1 음성구간 검출 알고리듬
Fig 4.1 speech interval detection algorithm

먼저 파일에 저장된 각 샘플 100개(10msec)씩을 한 블럭(block)으로하여 파일전체를 블럭화하고나서 식 (4.1)과같이 각 블럭의 구간에너지를 구한다.

$$E(n)=\sum_{m=1}^{100}|x(n,m)|^2, \qquad n=1,2,\cdots,N \qquad (4.1)$$

여기서 N은 블럭의 갯수, 즉 한 파일을 구성하고 있는 전체 샘플의 갯수를 100으로 나눈다음 이를 정수화한 값이다. $x(n,m)$는 제 n번째 블럭의 제 m번째 샘플값을 나타낸다.

이렇게하여 각 구간에서 에너지를 구한 다음, 이중에서 가장 큰 값 $E(p)$를 찾아내고, 그값의 5%를 넘는 에너지를 갖는 블럭을 파일의 앞과 뒤에서부터 찾아나가는 방법을 사용하는데, 이때 찾아진 두 블럭을 음성의 시작(n_S)과 끝(n_E)으로 하고 이 전후의 블럭들은 무음 구간이므로 제거해버리면 된다. 이때 5%라는 경계치는 많은 실험결과 경험으로부터 얻어진 수치다. 이 과정을 거치고나면 각 파일은 항상 100의 배수가 되는 샘플을 갖게되며, 파일의 크기가 60%정도로 감소한다.

이렇게하여 음성구간만을 걸러냈으므로 이제 각 파일에 대해서 특징성분을 추출한다. 이를 위해 먼저 각 음성신호에 해밍윈도우를 씌워 프레임화하는데, 이때 윈도우의 길이는 단어음성에 대해서 시간적·주파수적 분석을 위해 가장 적당한 32msec(320samples)로 결정하였다.

한 파일에 평균 33개의 프레임이 존재하는데, 이들 중에서 실험의 촛점에 따라 10개 내지 40개를 균일간격으로 뽑아내어 거기서 특징성분을 계산하는 것이다. 이는 전체 데이타의 30%에서 120%에 해당하는 양이며, 추출된 데이타량이 실제의 약 30%밖에 안되더라도 인접된 음성 샘플의 높은 상관성을 고려한다면 그다지 적은 양이 아니다. 또 프레임끼리 겹치는 비율을 음성파일의 길이에 따라 자동적으로 조절시킴으로써 실제보다 더 많은 분석 구간을 얻을 수 있게된다. 이로써 발성자의 발음지속시간이 각 파일마다 다른데에 따른 길이 변동의 효과를 상쇄시킬 수 있게된다. 사실 그동안 음성인식기술을 개발하는데에 있어서 음성의 길이 변동을 흡수하는 문제가 상당히 중요한 문제중의 하나였다. 음성 분석의 특징 성분으로 PARCOR계수를 사용했는데, AR(autoregressive) 모델에 기초한 자기상관법(autocorrelation method)을 사용하여 실험의 촛점에 따라 5차에서 20차의 PARCOR계수를 계산한다. 특징 성분 추출과정이 [그림 2]에 묘사되어 있다.

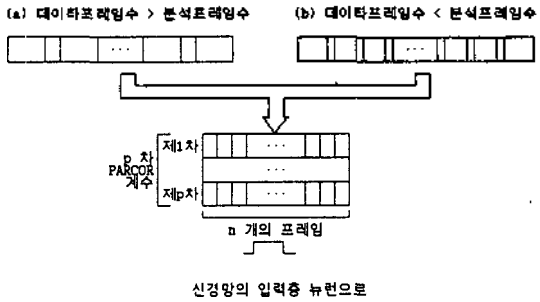

그림 2 특징 성분 추출 과정
Fig 2 process of extracting feature components

1명의 20대 남자가 하나의 단어를 10번씩 발음하였으므로, 10개의 데이타중에서 7개를 신경망의 학습데이타로 사용하고 나머지 3개의 데이타를 인식할 수 있는지 살펴본다. 따라서 학습용 데이타는 총 350개이고, 인식실험용 데이타는 총 150개다. 인공신경망의 학습계수(learning rate)에는 0.1을, 관성계수(momentum)에는 0.6을 주었으며, 신경망의 출력과 목표값의 차이, 즉 오차 E가 0.2 이내에 들 때 학습이 완료된 것으로 가정하였고, 만일 1500회를 반복하여도 위의 허용오차내에 들지 않을 경우 무조건 학습을 중지시켰다.

(1)전체학습과 구분학습

본 연구에서 실험 대상으로 삼고있는 단어의 종류는 모두 50종이며, 이들을 한꺼번에 신경망에 넣어 학습시키는 것에는 무리가 있을 것으로 보이므로 실험에 사용되는 대상 단어를 두 집단으로 나누어 각각 다른 인공신경망으로 학습시키는 것이 타당할지도 모른다. 이렇게 하면 하나의 신경망이 부담하는 학습량이 반으로 줄지만, 이와같은 방법에는 어떤 데이타가 어느 신경망으로 가야할지를 결정하는 또다른 신경망이 필요하게 된다. 이 제3의 신경망은 하나의 출력층 뉴런만 갖고 있으며 그 역할을 충분히 해낼 수 있는데, 그 출력이 0이면 제1집단을, 그리고 출력이 1이면 제2집단을 가리키는 것으로하면 그에 따른 뉴런간의 연결선, 즉 가중치의 갯수 또한 줄어들게되므로 오직 입력되는 데이타가 0이나 1이냐만을 판별하는 신경망으로서는 적당한 규모라고 볼 수 있다. 이런

방법의 개략도가 [그림 3]이다.

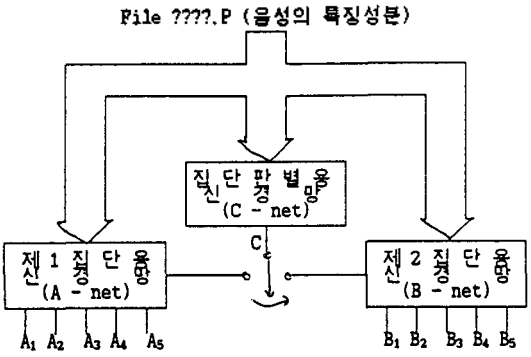

그림 3 구분 학습 방법
Fig 3 pre-classifying method

위와 같은 구분학습을 시키기 위해서는 50종의 단어를 두 집단으로 나누어주어야 할 필요가 있다. 이 구분의 기준은 언어의 음소중에 단어를 2분 또는 3분하는 무성음이 들어있는가 없는가인데, 일반적으로 무성음(unvoiced sounds)은 자기상관도(autocorrelation)가 유성음(voiced sounds)보다 낮기때문에 자기상관계산을 하면 단어의 무성음 포함여부를 가려낼 수 있게된다. 여기서는 자기상관도의 계산에 다음과 같은 식을 사용하였다.

$$\phi(n, k) = \sum_{m=1}^{N} x(m) \cdot w(m) \cdot x(m+k) \cdot w(m+k),$$

$$k = 0, 1, \cdots, M-1$$

$$r(n) = \max[\frac{\phi(n, k)}{\phi(n, 0)}], \; n = 1, 2, \cdots, N$$

여기서, w는 제2장에서 설명한 해밍윈도우이고, M은 한 프레임을 이루는 샘플의 갯수 (320)이며, N은 프레임의 갯수이다. 이때 프레임끼리는 10msec씩 겹쳐져 있다.

[그림 4]는 지명단어 /서울/의 음성파형과 자기상관도, 그리고 그것의 문턱가공(thresholding)된 결과를 보이고 있다. 음성의 전반부를 이루는 /ㅅ/ 음소부분에서는 자기상관도가 낮으나 /ㅓ울/ 부분이 모두 유성음으로 이루어져 있기때문에 자기상관도가 높다. $u(w\text{-}0.82)$는 [그림 4]의 (b)를 (c)로 변환하는 함수이며, 여기서 w는 각 프레임에서의 자기 상관값이다. 즉, 자기상관도가 0.82보다 큰 프레임에는 1을, 그렇지않은 프레임에는 0을 대입한다.

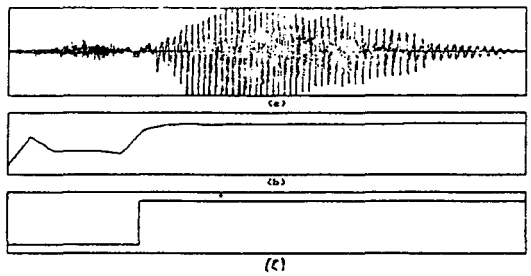

그림 4 /서울/에 대한 (a) 음성 파형 (b) 자기상관도(autocorrelation) (c) 문턱가공(thresholding)후
Fig 4 For /soul/(a) speech waveform (b) autocorrelation (c) after thresholding

[표 2]와 [표 3]에 나열된 바와 같이 제1집단은 22개의 지명 단어로 이루어져 있으며, 제2집단은 28개다. 여기서 유의해야 할 사항은 /ㄱ/이 보통은 무성파열음으로 분류되지만 [그림 5]와 같이 유성음사이에 끼여있을 때에는 유성음화하며, 그 사실이 그림에서 와같이 증명된다는 것이다.

표 2. 제1집단 단어군
Table 2. words of group 1

1. 서 울	6. 발 안	11. 외 왕	16. 원 당	21. 연무대
2. 대 구	7. 수 원	12. 하 남	17. 보 은	22. 동광양
3. 고 양	8. 성 남	13. 도 계	18. 영 월	
4. 광 명	9. 안 양	14. 설 악	19. 철 원	
5. 미 금	10. 용 인	15. 양 구	20. 태 백	

표 3. 제2집단 단어군
Table 3. words of group 2

1. 부 산	6. 화 천	11. 양 주	16. 파 주	21. 춘 천	26. 장호원
2. 광 주	7. 문 산	12. 여 주	17. 화 성	22. 평 창	27. 주문진
3. 대 전	8. 시 흥	13. 연 천	18. 원 주	23. 홍 천	28. 장승포
4. 강 화	9. 안 산	14. 오 산	19. 인 제	24. 남양주	
5. 과 천	10. 안 성	15. 일 산	20. 정 선	25. 의정부	

이와같이 자기상관값의 문턱가공결과 1의 구간이 하나인 지명 단어를 제1집단으로, 그리고 [그림 6]처

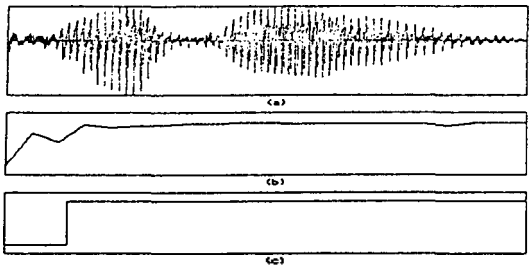

그림 5 /대구/에 대한 (a) 음성 파형 (b) 자기상관도(autocorrelation) (c) 문턱가공(thresholding)후

Fig 5 For/daegu/(a) speech waveform (b) autocorrelation (c) after thresholding

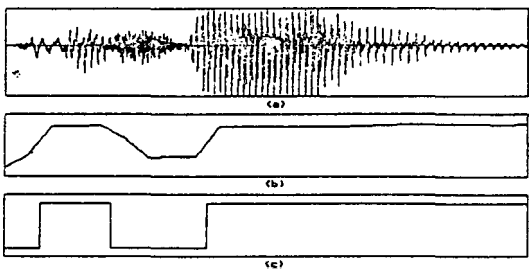

그림 6 /부산/에 대한 (a) 음성 파형 (b) 자기상관도(autocorrelation) (c) 문턱가공(thresholding)후

Fig 6 For/busan/(a) speech waveform (b) autocorrelation (c) after thresholding

럼 1의 구간이 둘 이상인 단어를 제2집단으로 분류하였다.

이와 같은 방법으로 대상 단어를 두 집단으로 나누어서 신경망 C-net, A-net, B-net을 각각 학습시켜서 얻는 인식률과 50개 대상 단어를 하나의 신경망으로 한꺼번에 학습시켜서 얻는 인식률을 비교조사하는 것이 첫번째 실험에 맞춰진 촛점이다.

이에는 4종의 시뮬레이션(simulation)이 포함되어 있으며, 20개 프레임에서 10차 PARCOR계수를 추출하였을 때 전체학습과 구분학습을 비교 실험하고, 또 15개 프레임에서 10차 PARCOR계수를 추출하였을 때 전체학습과 구분학습을 비교 실험한다. 전체학습의 경우 구분해야 할 단어수가 50개이며, 6개의 출력층 뉴런을 2진 코딩(binary coding)형식으로 구성하면 최대 64개까지 구분가능하므로 6개로 충분히 구분할 수 있다. 이와 같은 출력층 뉴런의 2진 코딩 형식이 또한 본 연구의 네번째 실험의 촛점이기도 하다. 네번째 실험이외에는 모든 출력층 뉴런을 2진 코딩 형태로 구성하기로 한다. 아울러 전체학습과 구분학습을 상호 비교하기위한 첫번째 실험이외에는 모두 구분학습법을 사용한다.

(2) 프레임갯수와 PARCOR차수

입력층 뉴런의 갯수는 단어에서 추출된 특징의 갯수와 같으므로 많을수록 그 단어를 확인하는 능력이 커질 것이 당연하다. 그러나 입력층 뉴런의 갯수를 한정없이 많이 만들수는 없다.

입력층 뉴런의 갯수는 단어에서 골라낸 프레임 갯수와 한 프레임내에서 계산된 PARCOR 차수를 곱한 것과 같으므로 위와같은 범위내에서 PARCOR 차수를 줄이는 대신 프레임 갯수를 늘리는 것이 더 높은 인식률을 가져올지, 아니면 PARCOR 차수의 중요도가 프레임 갯수보다 더 큰지를 알아보기위한 실험이 필요하다. 본 연구의 두번째 실험의 촛점은 바로 이것이다.

(3) 중간층 뉴런의 갯수

중간층에 분포하는 뉴런의 수에 따른 인식률의 변화를 관찰한다.

(4) 출력층 뉴런의 구성 방법

지금까지 전통적으로 인공신경망의 출력층 뉴런의 갯수는 구분해야 할 데이타의 종류수와 같았다. 따라서 구분해야 할 각각의 데이타마다 한개의 출력층 뉴런을 대응시켜서 한번에 한개씩의 출력층 뉴런이 활성화되도록 하였다. 이런 구성 방법이 신경망을 학습시키는데에 더 효과적이며 확실하지않은 신경망의 출력에도 입력된 데이타에 대한 후보의 수가 훨씬 적다는 것은 이미 알려진 바다.

그러나 구분해야 할 데이타가 아주 많은 경우에 구분해야 할 데이타수와 1대1 대응하는 출력층 뉴런을 만든다는 것은 신경망의 규모의 문제에 있어서 심각한 문제가 아닐 수 없다. 그에 따른 학습시간에 관련된 문제 또한 상당한 것이다.

본 연구의 네번째 실험에서는 50개의 단어를 구별하기 위해 구분학습방법을 사용하여 첫번째는 1대1 코딩방법으로 C-net에 2개, A-net에 2개, A-net에 22개, B-net에 28개의 출력층 뉴런을 두어 학습시켰고, 두번째는 2진 코딩(binary coding) 방법으로 C-net에 1개, A-net와 B-net에 각각 5개씩의 출력층 뉴런을 두어 학습시켜서 두 경우의 결과를 비교한다.

Ⅲ. 실험결과

(1)전체학습과 구분학습

[표 4]에서 횟수는 신경망이 학습을 완료하기까지 오차역전달 과정이 반복된 횟수를 말하며, 구분학습 행에서 합계는 C-net, A-net, B-net의 세 신경망이 학습을 완료한 데에 걸린 반복횟수와 시간을 모두 합한 것이며, 이때의 인식률은

$$\frac{C\text{-}net\text{의 인식률} \times (A\text{-}net\text{의 인식률} + B\text{-}net\text{의 인식률})}{2}$$

로서, A-net과 B-net의 인식률이 단독으로는 의미가 없고 C-net에서 집단 구분이 제대로 되어야만 비로서 그 의미를 가질 수 있으므로 (C-net의 인식률 × A-net의 인식률)과, (C-net의 인식률 × B-net의 인식률)의 평균값을 계산한 것이다. 각 신경망에서의 인식률은 학습된 신경망에 50종 × 3개의 음성데이타인 150개의 파일을 얼마나 제대로 분류할 수 있는지를 관찰한 결과이며, 150개중에서 15개의 데이타를 바로 분류하지 못했다면 인식률은 90%가 된다.

앞에서 볼 수 있듯이 전체학습시에는 두가지 실험조건 모두에서 최대 허용 반복횟수인 1500회의 반복을 통하고도 허용오차 0.2 내에 들지 못해 15시간 이상의 학습시간중에도 학습이 완료되지 않은 반면에, 구분학습시에는 학습완료까지 10시간을 넘지 않았음에도 불구하고 두 실험 모두에서 인식률이 전체학습

표 4. 전체학습과 구분학습의 실험결과

Table 4. experimental result of and pre-classifying method and non-pre-classifying

			횟수	시간	인식률(%)
프레임 갯수 20 PARCOR 차수 10 중간층뉴런수 20	전체학습		1500	19 : 19	78.7
	구분학습	C-net	69	0 : 32	98.7
		A-net	101	0 : 21	92.4
		B-net	1500	6 : 47	88.1
		합계	1669	7 : 30	89.1
프레임 갯수 15 PARCOR 차수 10 중간층뉴런수 20	전체학습		1500	16 : 28	68.0
	구분학습	C-net	55	0 : 20	94.0
		A-net	74	0 : 12	78.8
		B-net	103	0 : 22	82.0
		합계	231	0 : 54	75.6

때보다 더 높게 나타났다.

예상대로 전체학습법보다 구분학습법이 더 유리함을 실험을 통해 증명하였다.

(2) 프레임갯수와 PARCOR 차수에 대한 의존도

프레임의 갯수와 PARCOR 차수의 두가지 입력 조건중에 어느쪽이 인식률 향상에 더 큰 영향을 미치는지를 조사하는 실험이다. 결과는 [표 5]와 같다.

위로부터 세개의 실험 결과는 입력층뉴런이 200개,

표 5. 프레임갯수와 PARCOR차수에 대한 의존도

Table 5. Recognition rate dependency on the number of frames and the order of PARCOR coefficients

			C-net			A-net			B-net			합계		
프레임갯수	PARCOR차수	중간층뉴런수	횟수	시간	인식률(%)	횟수	시간	인식률(%)	횟수	시간	인식률(%)	횟수	시간	인식률(%)
10	20	20	78	0 : 36	95.3	57	0 : 12	86.4	703	3 : 09	92.5	838	3 : 57	85.2
20	10	20	69	0 : 32	98.7	101	0 : 21	92.4	1500	6 : 47	88.1	1670	7 : 30	89.1
40	5	20	89	0 : 42	99.3	89	0 : 19	80.3	424	1 : 55	84.5	602	2 : 56	81.8
10	15	50	71	0 : 51	94.0	68	0 : 22	72.7	55	0 : 23	86.9	194	1 : 36	75.0
15	10	50	58	0 : 42	94.0	64	0 : 21	80.3	57	0 : 24	82.1	179	1 : 27	76.3
30	5	50	70	0 : 51	94.0	138	0 : 46	78.8	64	0 : 26	77.4	272	2 : 03	73.4

중간층뉴런이 20개인 경우이고, 그 아래 세개의 실험결과는 입력층뉴런과 중간층뉴런이 각각 150개와 50개인 경우이며, 표에서 보는 바대로 입력층뉴런의 갯수가 많은 쪽의 인식률이 상대적으로 높다.

500개의 단어 음성의 길이가 제각기 다르나, 제4장에서 설명한 방법대로 분석을 위한 해밍윈도우를 서로 조금씩 겹치게 한 결과 평균적으로 한 음성 파일이 33개의 프레임을 갖고 있으며, 40개의 프레임을 분석용으로 추출한 경우에는 프레임간 중복이 그보다 더 커진다. 음성의 특징이 시간적으로 그리 빨리 변하지 않으므로 입력특징으로 200개의 성분을 가진 경우와 150개의 성분을 가진 두 경우 모두에 있어서 음성 파일에 대한 분석용 프레임의 총 길이가 50~60%인 경우에 대체적으로 그 음성의 정보를 거의 다 갖고있다.

따라서 40개의 프레임에서 5개씩의 PARCOR계수를 추출한 경우보다 20개의 프레임에서 10개씩의 PARCOR계수를 추출한 경우가 비슷한 양의 정보에 대해서 더 많은 수의 PARCOR계수를 추출한 꼴이 되어 인식률이 높게 나타나는 것이다. 같은 이유로 15프레임×10차 PARCOR계수를 사용한 실험이 30프레임×15차 PARCOR계수를 사용한 것보다 더 높은 인식률을 가져온다.

실험 결과 PARCOR차수가 5일때 단어 인식에 충분한 분석이 되지 못함을 볼 수 있다. 즉 자음 성분을 고려하면 한프레임에는 적어도 5차 이상의 분석을 해야된다. 그리고 PARCOR계수가 그보다 더 높아지더라도 분석을 위한 프레임의 갯수가 10프레임정도로 너무 적으면(약 30%) 결과로서의 인식률은 더 높아지지 않는다.

(3)중간층 뉴런의 갯수

다른 모든 조건을 동일하게 했을 때 중간층 뉴런의 수가 인식률에 어떤 영향을 미칠 것인가에 대한 실험결과다.

표에서 보듯이 중간층뉴런의 갯수가 턱없이 적으면 전혀 패턴 분류가 이루어지지 않는다. 30프레임×5차 PARCOR계수 실험에서, 그리고 15개 프레임×10차 PARCOR계수실험에서 신경망의 학습시 반복 횟수가 모두 1500회로 나타나 있으며, 이는 허용반복 횟수내에 신경망이 학습되지 않았음을 말해준다.

중간층뉴런이 20개인 경우와 50개인 경우에 대한 실험 결과에서는 인식률의 변동 및 학습에 요구되는 시간의 차이가 그다지 크게 느껴지지 않는다. 따라서 위의 실험결과로부터 중간층 뉴런의 갯수가 20개에서 50개로 늘어나는 정도로는 학습및 인식률에 그다지 큰 영향을 미치지 않는다는 결론을 내릴 수 있다.

중간층뉴런의 갯수의 영향에 대한 좀 더 넓은 결론을 얻기위해서는 뉴런의 갯수를 100이상으로 높여보는 실험이 필요할 것이나, 제4장에서 언급한 바대로 실험환경의 제약조건에 따라 이에 관한 실험은 현재 보류중이다.

(4)출력층 뉴런의 구성 방법에 따른 변동

각 패턴에 대해 전용의 출력 뉴런을 하나씩 두고 한 패턴에 대해서 하나의 뉴런만이 활성화되도록 하

표 6. 중간층뉴런의 갯수에 대한 의존도
Table 6. Recognition rate dependency on the number of neurons in the hidden layer

		C-net			A-net			B-net			합 계		
입력 패턴	중간층뉴런수	횟수	시간	인식률%	횟수	시간	인식률%	횟수	시간	인식률%	횟수	시간	인식률%
프레임 갯수 30 PARCOR 차수 5	5	1500	4:46	48.0	1500	2:08	0	1500	2:44	0	4500	9:38	0
	20	106	0:40	94.0	353	1:00	77.3	106	0:23	89.3	565	2:03	78.3
	50	70	0:51	94.0	138	0:46	78.8	64	0:26	77.4	272	2:03	73.4
프레임 갯수 15 PARCOR 차수 10	5	1500	4:42	50.0	1500	2:08	0	1500	2:42	0	4500	9:32	0
	20	55	0:20	94.0	74	0:12	78.8	103	0:22	82.1	232	0:54	75.6
	50	58	0:42	94.0	64	0:21	80.3	57	0:24	82.1	179	1:27	76.3

는 방법과 동시에 두개 이상의 뉴런이 활성화되는 것을 허용하여 각 패턴에 일련번호를 붙이고 그에 대한 이진수(binary) 형태로 출력을 구성하는 방법의 두 가지를 생각해 볼 수 있다. 이에 대한 실험 결과가 [표 7]이다.

입력패턴에 대한 1대1 구성을 하면 특정 패턴에 대해 하나의 출력 뉴런에 연결된 가중치만이 집중적으로 학습되며, 학습 완료된 후에도 어떤 패턴에 대한 출력 뉴런의 활성도가 기준치 이하가 되더라고 기준치를 다시 조정해보면 그 해(解)를 찾을 수 있는 경우가 많다. 그러나 출력 뉴런이 동시에 여러개 활성화되도록하면 학습의 집중도가 분산될 뿐만 아니라 학습 완료후에도 앞에서와 같은 문제가 생겼을 때 기준치를 재조정해도 해가 두개 이상 생기기 때문에 패턴 분류율이 대체적으로 낮은 것이 당연하다. 그러나 실험 결과에서처럼 그 차이는 20%미만인 반면에 학습에 걸리는 시간은 50% 이상 감소되며, 적은 수의 출력 뉴런을 갖고 많은 패턴을 분류할 수 있으므로 나름대로 그 의미를 갖는다고 하겠다.

1대1 대응의 출력 구성시에 표의 내용과 같이 제한시간 내에 학습되지 않았으며, 이는 신경망이 국부계곡(local minimum)에 빠진 것으로 보이는데, 실제로 15프레임 × 10차 PARCOR계수에 대한 1대1 대응 실험에서 B-net는 패턴/부산/과 /강화/가 전혀 학습되지 않아 28개의 출력 뉴런이 세번의 인식 기회에 모두 0의 출력을 보였다.

한 실험이 다른 비교에도 사용된 것을 제외하면 총 14개의 시뮬레이션이 본 실험에서 행해졌는데, 구분학습법을 사용했을 때의 인식률이 전체학습법에 비해 높으며, 200개의 패턴 특징(10개 프레임 × 10차 PARCOR계수)을 사용한 경우에도 2진 코딩 형식의 출력을 구성한 쪽이 150개의 패턴 특징(15개 프레임 × 10차 PARCOR계수)을 사용하더라도 1대1 대응의 출력 형식을 사용한 쪽보다 89.1%대 89.6%로 인식률이 떨어진다. 이로써 입력층 뉴런의 갯수의 많고 적음보다 출력층 뉴런의 출력 형식이 인식률이 더 큰 영향을 미친다는 것 또한 알 수 있게된다.

Ⅳ. 결 론

본 연구에서는 2, 3음절의 단어를 인공신경망을 이용하여 인식하는 실험을 하였다. 50개의 대상단어를 한꺼번에, 그리고 두 집단으로 나누어서 구별하도록 한 결과 비록 제한적이기는 하지만 구분학습법을 사용한 인식결과가 더 좋았으며, 32msec의 길이를 갖는 프레임의 갯수는 음성 전체 길이의 50%~60% 정도를 분석 구간으로 택할 수 있도록 한 경우가 가장 적당함으로 알 수 있었다.

표 7. 출력층 뉴런의 형식에 따른 변동

Table 7. Recognition rate dependency on the composition method of output layer

			C-net			A-net			B-net			합 계		
입력 패턴	중간층 뉴런 갯수	출력층뉴런구성형식	횟수	시간	인식률%	횟수	시간	인식률%	횟수	시간	인식률%	횟수	시간	인식률%
프레임 갯수 15	20	1대1 대응	60	0 : 22	96.7	716	2 : 17	98.5	1500	6 : 21	86.9	2277	9 : 00	89.6
PARCOR 차수 10	20	2진 코딩	55	0 : 20	94.0	74	0 : 12	78.8	103	0 : 22	82.1	232	0 : 54	75.6
프레임 갯수 30	20	1대1 대응	75	0 : 28	92.7	1500	4 : 48	90.9	1500	6 : 22	92.9	3075	11 : 38	85.2
PARCOR 차수 5	20	2진 코딩	106	0 : 40	94.0	353	1 : 00	77.3	106	0 : 23	89.3	565	2 : 03	78.3

입력과 중간층의 노드이 수가 같은 상황에서 출력층 뉴런의 출력 형식은 2진(binary) 코딩 방법을 사용하는 것보다 1대1 대응방법을 사용하는 것이 더 높은 인식률을 가져오는 것이 사실이지만, 2진 코딩 방법쪽이 학습 시간 단축 효과를 줄 뿐만 아니라 어휘의 양이 많아질 때 그에 따라 출력 뉴런 수가 선형적으로 증가하지 않는 장점이 있다..

실험에 사용된 14종의 시뮬레이션 결과, 15프레임에서 10개의 PARCOR계수를 추출하여 150개의 입력층 뉴런과 20개의 중간층 뉴런을 통해 1대1 대응 형식의 출력층 뉴런으로 인식시키는 방법이 89.6%로 가장 높은 인식률을 보였다. 출력층 뉴런의 출력을 2진 코딩 형식으로 구성한다면 20개 프레임×10차 PARCOR계수를 사용한 결과가 89.1%로 가장 높다.

본 연구는 대상 단어에 대한 음소별 또는 음절별 분석을 수행하지 않고 전체적으로 처리하는 방법에 관한 것이기 때문에 신경망이 인식해야 할 단어를 모두 미리 학습하고 있어야 하며, 그 어휘의 규모가 그다지 크지 않다면 인식률을 조금 더 향상시켜서 실용화되게 할 수 있을 것이다.

참 고 문 헌

1. 김석동·이행세 *신경망을 이용한 우리말 인식에 관한 연구* 1992 *한국음향학회지* 제11권 제3호 pp.14-24
2. 임택규·이재건·김석동·이행세 *인공신경망을 이용한 대표종성음 인식방법* 1992 *대한전자공학회 하계종합학술대회논문집 Vol.15 No.1 pp.713-717*
3. *이재건·임택규·김선일·김석동·이행세 음소인식에 의한 한국어 단음절 인식* 1992 *대한전자공학회 추계종합학술대회논문집* Vol.15 No.2 pp.665-669
4. 김석동·이행세 *복합신경망을 이용한 우리말 단음인식에 관한 연구* 1992 *한국음향학회지* 제11권 제6호 pp. 23-31
5. 이욱재·이재건·송도선·이행세 *우리말 복모음인식에 관한 연구* 1993 *제3회 인공지능, 신경망 및 퍼지시스템 종합학술대회논문집* pp.217-219
6. 심성룡·이욱재·이재건·김석동·이행세 *BP알고리듬을 사용한 한국지명음성의 인식방법* 1993 *대한전자공학회 추계종합학술대회논문집* Vol.16 No.2 pp.992-995
7. 오영환 *패턴인식론* 1991 정익사
8. 안수길 *음성분석, 모델링* 1993.5 *대한전자공학회지* 제20권 제5호 pp.62-70
9. R.D.Peacoke and D.H.Graf *An Introduction to Speech and Speaker Recognition Computer* Vol.23 No.8 August 1990
10. W. Jones and J.Hoskins *Back-Propagation Byte* Oct. 1990
11. R.P.Lippmann *Review of Neural Networks for Speech Recognition Readings in Speech Recognition* 1990 Morgan Kaufman, Inc
12. L.R.Rabiner and R.W.Schafer *Digital Processing of Speech Signals* 1978 Prentice-Hall Inc.
13. S.Saito and K.Nakata *Fundamentals of Speech Signal Processing* 1985 Academic Press
14. T.W.Parsons *Voice and Speech Processing* 1986 McGraw Hill Inc.
15. S.Furui *Digital Speech Processing, Synthesis, and Recognition* 1992 Marcel Dekker Inc.
16. J.R.Deller Jr., J.G.Proakis and J.H.L.Hansen *Discrete-Time Processing of Speech Signals* 1993 McMillan Co.
17. J.M.Zurada *Introduction to Artificial Neural Systems* 1992 West Publishing Co.
18. J. L.McClelland, D.E.Rumelhart and the PDP research Group *Parallel Distributed Processing* Vol.1, 2 1986 MIT Press
19. P.D.Wasserman *Neural Computing: Theory and Practice* 1989 Van Nostrand Reinhold New York
20. J.A.Freeman and D.M.Skapura *Neural Networks: Algorithms, Applications, and Programming Techniques* 1991 Addison-Wesley Publishing Co.
21. M.Caudill and C.Butler *Understanding Neural Networks: Computer Exploration* Vol.1, 2 1992 MIT Press
22. B.Kosko *Neural Networks and Fuzzy Systems* 1992 Prentice-Hall Inc.

▲송 도 선 : 제12권 4호 참조
▲김 석 동 : 제12권 4호 참조
▲이 행 세 : 제12권 4호 참조