통합 사용자 인터페이스에 관한 연구 :
인공 신경망 모델을 이용한 한국어 단모음 인식 및 음성 인지 실험

이 봉규*, 김 인범, 김 기석, 황 희융
서울대학교 컴퓨터공학과

A Study on the Intelligent Man-Machine Interface System : The Experiments of the Recognition of Korean Monotongs and Cognitive Phenomena of Korean Speech Recognition Using Artificial Neural Net Models.

Lee Bongku, Kim Inbum, Kim Kiseok, Hwang heeyeung
Dept. of Computer Engineering, Seoul National University

(요 약)

음성 및 문자를 통한 컴퓨터와의 정보 교환을 위한 통합 사용자 인터페이스(Intelligent Man- Machine interface)시스템의 일환으로 한국어 단모음의 인식을 위한 시스템을 인공 신경망 모델을 사용하여 구현하였으며 인식시스템의 상위 접속부에 필요한 단어 인식 모듈에 있어서의 인지 실험도 행하였다. 모음인식의 입력으로는 제1, 제2, 제3 포르만트가 사용되었으며 실험대상은 한국어의 [아,어,오,우,으,이,애,에]의 8개의 단모음으로 하였다. 사용한 인공 신경망 모델은 Multilayer Perceptron이며, 학습 규칙은 Generalized Delta Rule이다. 1인의 남성 화자에 대하여 약 94%의 인식율을 나타내었다. 그리고 음성 인식시의 인지 현상 실험을 위하여 약 20개의 단어를 인공신경망의 어휘레벨에 저장하여 음성의 왜곡, 인지시의 lexical 영향, categorical percetion등을 실험하였다. 이때의 인공 신경망 모델은 Interactive Activation and Competition Model을 사용하였으며, 음성 입력으로는 가상의 음성 피쳐 데이타를 사용하였다.

I.서론 : 한국어 음성인식 시스템

음성인식 시스템은 음성신호로부터 음성인식의 최소단위(일반적으로 음소)로 transform하는 과정과, 그 최소 단위의 string으로 부터 단어, 의미, 문법적 지식을 동해 문장을 만들어 내거나, 그 과정에서 하부과정을 다시 access하는 등의 두 단계로 나뉘어진다. 전 단계에서는 음성 신호의 sampling, signal transform, LPC 또는 FFT추출, 특징의 추출및 음소의 해석등의 작업이 있으며 상위 과정에서는 단어 지식, 의미 및 문법 지식을 동하여 부정확한 음소 string의 보완 및 해석 작업이 이루어 진다. 이러한 음성패턴의 인식의 과정에는 많은 방법이 있을 수 있으나, 본 논문은 인간의 인지 모델의 원리를 따르는 인공 신경망 모델에 근거하여 연구를 수행하였다. 그리하여 일단 음성신호로 부터 모음부의 추출및 분류작업을 인공 신경망 모델로 설계하여 실험하였으며, 그 결과를 근거로 앞으로 음성신호-음소 인식부를 설계할 예정이다. 그리고 상부과정의 실험을 단어지식 선에서만 행하였다. 이를 통하여 인간의 음성 인지 과정을 어떻게 인공 신경망 모델을 통해 이를 수 있는지를 실험해본다.

II. 한국어 모음 인식 시스템

1. 모음 분류의 필요성

모음은 자음보다는 정보 전달량은 작다. 그러나 음성 인식 시스템의 구현에 있어 자음의 분류보다는 모음부의 분류가 더 쉽고 안정적이다. 따라서 자음부의 분류 이전에 먼저 모음부를 찾아내어 안정점을 찾은뒤 이 정보를 중심으로 자음부의 분류가 이루어지는 것이 일반적인 음성 인식의 절차이다. 따라서 한국어 음소 인식 시스템의 구현을 위해서도 음성에서의 모음부를 찾아내기 위한 알고리즘의 개발이 필요하며 모음부의 분류를 위한 실험 및 이를 위한 정확한 지식을 찾아내는 연구가 필요하다. 또한 한국어는 영어나 일어에 비해 모음의 존재가 뚜렷하고 모음의 구조가 복잡하며 무성음 뒤에는 반드시 모음이 온다는 특징이 있다. 따라서 한국어 음소 인식 시스템의 구현시 모음의 정확한 분류 및 인식은 중요한 비중을 차지하게 될것이다[1].

2. 한국어 모음의 종류 및 특성

음성은 크게 모음과 자음으로 나뉘어진다. 모음은 발음할 때에 성대에서 진동된 공기가 입 안에서 아무런 장애도 받지 않고 자유로이 입밖으로 흘러나오는 소리이다. 이와 반해 자음은 입안에서 여러가지 장애를 받아 발음되는 소리이다. 모음은 단순모음과 이중모음으로 나뉠 수 있다. 단순모음은 처음부터 끝까지 소리값에 변화가 없이 내는 /이/, /아/, /어/같은 모음을 말한다. 이중모음은 처음과 나중의 소리값이 달라지는 모음이며 /와/, /야/, /의/와 같은 모음이 여기에 속한다. 다음은 한국어의 모음을 단모음,이중모음,또는 반모음으로 분류해 놓은 것이다. [3]

단모음 : [이,에,애,아,어,오,우,으,외]
이중모음 : [와,야,여,요,유,의,위]
반모음 : [j] : [예,애,야,여,요,유]
[w] : [위,외,왜,와,워]

% 이 논문은 1989년도 문교부 지원 한국 학술 진흥 재단의 자유 공모 연구 과제 학술 조성비에 의하여 연구 되었음.

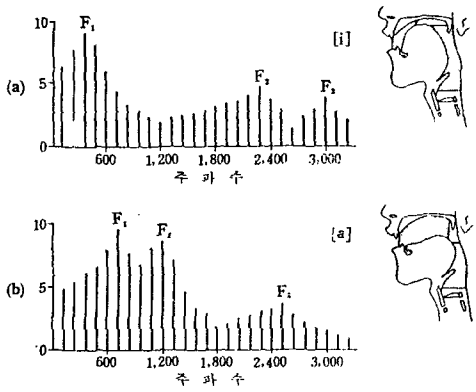


그림 1. [i]와 [a]의 스펙트럼과 안면도

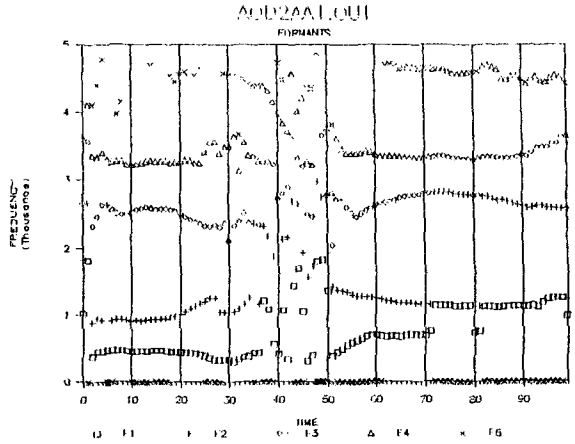


그림 2. / 어따 / 음성의 포르만트 값

3. 모음 인식을 위한 특징 파라미터

모음의 자동 분류 인식을 위해 가장 많이 사용되는 파라미터로는 포르만트가 있다. 포르만트(formant)란 어떤 한 음성의 음가가 다른 딴 음성과 식별되게 하여 하나의 독립된 음가를 지니게 하는 공명강의 진동 주파수 성분을 말한다. 이 포르만트는 음성기관의 주요한 공명강인 인강과 구강의 공명에서 나오는 것으로 알려져 있으며 이 값은 혀와 입술의 움직임에도 관계한다. 모음의 포르만트는 후두소리와 성도 전체가 가져오는 공명실의 영향,그리고 조음컴의 종합 등이 모두어져서 되는 것이다.[1]

인간의 성대(Vocal band)는 클라리넷의 리드와 마찬가지로 폭넓은 주파수대의 소리를 낸다. 이 소리에 대해 구강,비강,인강은 공명체 구실을 하게되며 이들 공명체의 여러 다른 모양의 조합이 여러가지의 공명곡선을 만들어 성대에서 나오는 소리 가운데서 이 공명 곡선에 맞는 소리만이 공명 하게 된다. 즉 여러 다른 모음들은 구강과 인강의 모양을 여러가지로 달리 함으로써 만들어지게 된다. 이 공명은 성도의 공명 곡선의 특성에 따라 이루어지는데 대략 기본 주파수 500c/s,1500c/s,2500c/s가 봉우리를 이루고 있으며 이들을 중심으로 소리의 에너지가 집중되어 있다. 이때의 에너지가 집중된 부분의 주파수를 포르만트라 한다. 그림 1에는 [이]와 [아]의 제 1,2,3 포르만트를 나타내는 스펙트럼이 나타나 있다. 이 그림을 통하여 우리는 포르만트가 어떻게 모음의 분류에 중요한 특징 변수가 되는지를 알수있다. [이]의 경우에는 혀의 앞부분이 매우 높아서 뒷 부분이 매우 길고 앞 부분이 매우 짧다. 그 결과 F1이 비교적 낮고(280c/s) F2가 높다(2240c/s). 여기에 비해 [아]를 발음할때의 성도의 모양은 [이]와 아주 다르다. [이]의 경우와는 반대로 뒷 부분이 짧아짐으로써 F1이 높아지고(710c/s) 앞 부분이 길어 짐으로써 F2가 [이]의 경우보다 낮다(1100c/s). 이것으로 보아 F1은 성도의 뒷부분과, F2는 성도의 앞부분과 관계 있음을 알수있다. 공기의 체적이 커질수록 공명 주파수가 낮아진다. 즉 F1의 주파수가 높아지는 것은 성도 뒷부분의 용적이 작아지는 것을 의미한다. 반대로 F2의 주파수가 낮아진다는 것은 앞부분의 용적이 커진다는 것을 의미한다.

포르만트는 스펙트로그램을 통하여 관찰할 수 있다. 스펙트로그램은 한 순간의 스펙트럼의 특성을 시간의 흐름에 따라 보여주는 기계이다. 뿐만 아니라 컴퓨터 프로그램에 의하여 포르만트의 값을 추출할 수 있다. 그림 2에는 한국어 음성 /어따/를 컴퓨터 프로그램에 의하여 포르만트를 구한 것을 LOTUS 1_2_3을 이용하여 나타낸 것이다. 이 그림으로부터 /어/에 해당하는 포르만트의 값과 /아/에 해당하는 포르만트의 값이 서로 약간의 차이가 있음을 나타낸다. 따라서 포르만트 값의 분포로 모음의 식별을 할 수 있다. 그림 3에는 영어 모음의 제 1, 제 2 포르만트의 분포를 나타냈다. 그림에서 보면 제 1, 제 2 포르만트의 값의 분포를 통하여 어느 정도 모음의 분류가 이루어질 수 있음을 볼 수 있다.

4. 포르만트를 이용한 모음인식 실험

위와 같이 포르만트는 모음의 식별을 위한 중요한 단서가 되어 있다. 따라서 모음의 분류를 위해 포르만트를 사용한 연구들이 많이 발표되어 있다. 그러나 이들 연구의 많은 한계점은 주어진 모음 집합들을 포르만트가 완전히 구별해주지는 않는다는 점이다. 그림 3에서 보는 바와 같이 즉 다른 모음과 겹쳐지는 영역들이 존재하며 또한 이들의 분포는 잡음 및 포르만트 추출 알고리즘의 부정확성, 발음하는 화자의 발음 습관의 차이, 성도의 크기의 차이, 기본 주파수의 차이등의 이유 때문에 일정치 못하다. 따라서 기존의 수학적 모델에 의한 확률적 분포 계산으로 주어진 도메인에서의 적절한 영역의 구분에는 한계가 있다.

따라서 이들을 구별해 줄 수 있는 적응력있는 모델이 필요하며 바로 인공 신경망 모델이 그러한 적응력이 매우 높은 것으로 나타났다. 인공신경망을 이용하여 모음의 분류를 행한 실험들은 최근에 일부가 진행되고 있는 중이다. 그중에서 포르만트를 이용하는 대표적인 것은 MIT의 Lippmann과 Huang이 실험한 것이 있다. 이 실험은 다층 구조의 인공 신경망을 사용하여 피터슨과 바니가 얻어놓은 영어 10개의 모음에 대한 남성,여성,어린이의 발음에 대한 포르만트1,포르만트2 를 이용하여 이들 모음을 인식하는 실험을 해보았다. 이 실험에서 사용한 인공 신경망의 구조는 3개의 층을 가지고 있는데 각 층은 입력 층이 200개의 unit으로, 히든 층은 80개의 unit으로,출력층은 10개의 unit으로 구성되어 총 290개의 unit으로 구성되어 있다. 이 실험에서는 약 80%의 인식율을 보였다. 본 논문에서는 이러한 연구를 한국 모음을 대상으로 실험해 본다.

III. GDR(Generalized Delta Rule)

1. 개요

1950년대 말에 인간의 신경 구조를 모방한 아주 간단한 모형인 퍼셉트론(Perceptron) 이라는 것이 미국의 Rosenbratt에 의하여 만들어졌다. 퍼셉트론을 가지고서 많은 사람들이 실험을 한 결과 어떤 문제에 대해서는 아주 잘 적용되어질 수 있음이 증명되었다. 그러나 그 후에

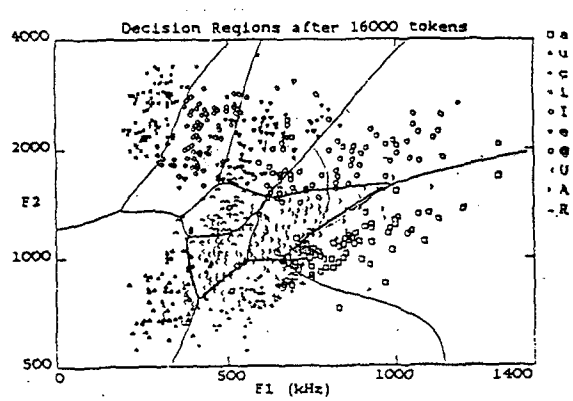


그림 3. 영어 모음의 F1,F2 분포

Minsky와 Papert는 "Perceptron(1969)"이라는 저서를 통해 퍼셉트론의 한계를 지적하였는데 그 대표적인 예가 exclusive or(xor) 문제이다. 즉 퍼셉트론을 가지고는 xor문제를 해결할 수 없음이 밝혀졌다. 이 후로 많은 학자들이 이 방면의 연구를 포기하고 소수의 학자들만이 연구를 계속해 왔었다. 그러던 중 1980년대에 들어서면서 D.Rumelhrat, J.MaClelland등을 주축으로 하는 일련의 연구 그룹이 형성되어 '병렬 분산 처리(Parallel distributed processing)'이라는 슬로건을 내걸고 많은 연구가 다시 진행되기 시작했다. 이 그룹에서 퍼셉트론의 한계로 지적되어진 문제에 대한 해결책들을 제시하였다.

이러한 해결책으로 제시된 것의 하나가 바로 인공 신경망의 구조를 다층화하여 입출력 층안에 내부 층을 두어 문제를 해결하는 모델(Multilayer Perceptron)으로 이는 Minsky에 의하여도 생각되어졌으나 그 당시에는 내부 층을 학습시킬 수 있는 적절한 학습 방법의 부재가 문제였었다. PDP그룹에서는 내부 층을 학습시키는데 유용한 학습 법칙을 고안했는데 이것이 바로 GDR(Generlized Delta Rule)이다.

2. 학습 규칙

GDR은 다름 아닌 단층 구조의 인공 신경망의 유용한 학습 법칙으로 알려진 델타 법칙(Delta rule) 의 일반화이다. 우선 간단히 델타 법칙에 관하여 알아보기로 하자.

1) 델타 법칙(Delta rule)

델타 법칙은 연결된 두 노드의 연결 강도의 변화량에 대한 법칙으로써 다음과 같이 나타낼 수 있다.

$$\Delta_p w_{ji} = \eta (t_{pj} - o_{pj})\, i_{pi} = \eta \delta_{pj} i_{pi}$$

여기서$\Delta_p w_{ji}$는 노드 i에서 노드 j로의 연결 강도(weight)를 표시하고 η은 학습 비율 (learning rate),t_{pj}는 노드 j의 원하는 출력 값, o_{pj}는 노드 j의 실제 출력 값, i_{pi}는 노드 i의 활성화 값을 나타낸다. 이 법칙은 각각의 학습 주기마다 $t_{pj} - o_{pj}$의 값을 구한 다음 입력의 활성화 정도를 곱한 값을 입출력 노드간의 연결 강도를 보정해 주는 값으로 사용하는 것으로서, 보정된 연결 강도는 전체 시스템의 에러를 최소화 하는 값이다. 이를 좀더 자세히 살펴보면 우선 전체 시스템의 에러는 다음과 같이 주어진다.

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

여기서 p는 입력 패턴의 갯수를 가리킨다. 이러한 전체 시스템의 에러를 최소화하는 방법의 하나로서 델타 법칙은 gradient-descent 방법을 이용하는데 gradient는 다음과 같이 계산되어진다.

$$\Delta_p w_{ji} \propto -\frac{\partial E_p}{\partial w_{ji}},$$

이 식을 풀어 보면

$$\Delta_p w_{ji} = \eta \delta_{pj} o_{pi}$$

되어 델타 법칙과 같음을 알 수 있다.

2) 일반화된 델타 법칙(Generalized Delta Rule)

다층 구조의 신경망에서의 유용한 성질은 입출력 계층간의 내부 노드의 존재 및 노드의 활성화 함수(activation function)의 비선형성에 기인한다. 선형 활성 함수를 가지는 다층 구조의 신경망은 같은 성질을 만족하는 단층 구조의 신경망으로 구현되어질 수 있음은 쉽게 증명되어 질 수 있다. 비선형 활성 함수를 가지는 다층 구조 신경망의 학습 법칙은 다음과 같다. 한 노드의 입력과 출력은 다음과 같이 정의되어 진다.

$$net_{pj} = \sum_i w_{ji} o_{pi},$$

$$o_{pj} = f_j(net_{pj})$$

여기서 f가 바로 비선형 활성 함수인데 이는 미분 가능해야 하고 증가 함수이어야 한다. 이러한 함수로 많이 쓰이는 것이 logistic 함수로서 다음과 같다.

$$f'_j(net_{pj}) = \frac{1}{1 + e^{-(\sum_i w_{ji} o_{pi} + \theta_j)}}$$

다층 구조 신경망에서 연결 강도의 변화량을 다음과 같이 가정한다.

$$-\frac{\partial E_p}{\partial w_{ji}} = \Delta_p w_{ji}$$

이를 풀면 다음과 같이 된다.

$$\Delta_p w_{ji} = \eta \delta_{pj} o_{pi}$$

이는 앞서 살펴보았던 델타 법칙과 동일함을 알 수 있다. 그러나 일반화된 델타 법칙에서는 δ_{pj} 의 계산이 노드 j가 출력 노드일 경우와 내부 노드일 경우에 다르게 계산된다. 즉 노드 j가 출력 노드라면 δ_{pj}는 다음과 같이 주어지고

$$\delta_{pj} = (t_{pj} - o_{pj}) f'_j(net_{pj})$$

노드 j가 내부 노드일 경우 δ_{pj}는 재귀적으로 다음과 같이 구해질 수 있다.

$$\delta_{pj} = f'_j(net_{pj}) \sum_k \delta_{pk} w_{kj}$$

이 식을 설명하면 내부 노드에서의 δ_{pj}는 바로 윗 노드에서 계산된 δ_{pk}를 이용하여 구해지는데 이 값은 노드 j로 들어오는 입력의 활성 함수의 미분 값,즉 활성의 변화량에 비례하여 조정함을 알 수 있다. 이러한 방법으로 에러를 아래 노드로 전달하면서 시스템의 에러를 최소화하는 방법이 일반화된 델타 법칙이다.

IV.MLP에 의한 모음 인식 실험과 실험 결과

1. 데이타의 획득

본 연구를 위해 사용되어진 음성 데이타로는 한국어에서 가능한 모음-자음-모음(VCV) 연쇄 음성을 사용하였다. 대상 모음으로는 한국어의 단모음중 많이 사용되는 8개 즉 V = [아,어,오,우,으,이,에,애]로 하였으며 자음으로는 파열음 중에서 몇개를 선택하였다.

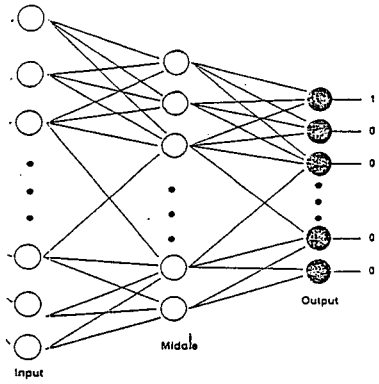
음성 데이타를 얻기 위한 녹음 환경으로는 일상적인 잡음이 들어갈 수 있는 조용한 방이다. microphone을 통하여 발성된 음성을 증폭기를 통하여 메탈테이프에 녹음된다. 녹음된 음성은 다시 증폭기를 통하여 적절한 크기로 증폭된 뒤 저역 통과기(LPF)에 들어간다. LPF의 차단 주파수는 4.95KHz이며 동시에 100Hz 미만의 주파수 대역이 차단된다. A/D Converter에서의 sampling frequency는 10KHz, 양자화 비트 수는 신호를 포함하여 12 bit 디지탈 신호로 변환한다.

컴퓨터에 저장된 음성 신호는 ILS-PC를 사용하여 음성 부분만 구분되어 1.2M floppy diskette에 저장된다. 이때 한 음성 데이타의 크기는 57478 byte이며 한 프레임의 point수를 64 point로 하였을때 120 frame의 크기이다.

이를 컴퓨터가 다시 읽어 처리할 때에 한 frame length는 256 sample (25.6ms), frame interval 64 sample(6.4ms)로 분석되었으며,hamming window를 사용하였다.

2. 인공 신경망 모델의 구조

본 논문에서 제안하는 모음의 분류를 위한 인공 신경망 모델의 구조는 다층구조(multi-layer)로 하였다. 여기서 다층구조라 하는 것은 인공 신경망이 입력 층(input-layer)과 출력 층(output-layer), 그리고 이들 층사이에 히든 층(hidden-layer)이라는 내부의 층을 하나 더 가지고 있다는 것이다. 즉 3개의 층을 가지고 있는 계층적인 구조라는 것이다. 이들 각각의 층은 서로 인접해 있는 층과 완전 연결(fully-connected)이 되어있고 이들 연결(connection)에는 고유의 가중치(weight)가 있어서 이들 가중치를 변경함으로써 인공 신경망의 학습은 이루어진다. 처리의 흐름은 입력 층이 외부로부터 입력을 받아들이고 이것의 처리 결과를 히든 층에 전하고 히든 층은 다시 결과를 출력 층에 전하게 된다. 그리고 이 신경망에서는 처리의 피드 백(feedback)이 존재하지 않으며 입력 층에서 출력 층으로 혹은 출력 층에서의 입력 층으로의 처리의 흐름은 없다. 각 계층에서의 unit들은 하위의 층에서 들어오는 입력을 적당한 함수(function)으로 처리하여 결과를 낸 다음 이 결과를 상위의 층으로 전달한다.따라서 하나의 unit들의 기능은 간단한 함수의 처리만을 하게 된다. 모음 인식을 위한 인공 신경망 모델(Multi-layer Perceptron)의 구조와 학습 사양이 그림 4에 나타나 있다.



unit 수	가변
input unit	40
output unit	8
hidden unit 1	30
hidden unit 2	가변

그림 4. 모음 인식을 위한
Multi-layer Perceptron
(a) 구조 (b) 사양

각 층의 기능을 상세히 살펴 보면 입력 층에서는 음성 데이타를 읽어서 이 음성 데이타 중 모음을 나타내는 부분의 중심을 찾은 다음 이 중심에서 좌 우로 한 프레임(frame)씩 포르만트 1,2,3을 구하여 이것을 이용하여 입력벡타를 만든 다음에 이 만들어진 입력 벡타를 인공 신경망 모델의 입력으로 읽어들이게 된다. 출력 층에서는 실제로 인공 신경망에서 인식된 모음이 8개 중에서 어디에 속하는 음이라는 것을 출력으로 표현한다. 따라서 출력 층의 unit수는 각 모음 당 하나 씩으로 정의하여 8개를 사용한다. 출력 층에서의 인식된 모음이 무엇인가를 표현하는 방법은 인식되어진 결과만이 1에 가깝게 되고 나머지는 모두 0에 가까운 수를 나타냄으로써 인식됨을 표현한다.

입력 층과 출력 층 사이에 존재하고 있는 히든(hidden)층은 아직까지 그 역할이 정의되어 있지 않고 있지만 지금까지 알려진 바로는 대략 들어오는 입력 패턴을 출력 패턴과 매칭시키기 위한 내부의 표현으로 변환한다고 한다. 그리고 이번과 같은 음성인식에서는 이러한 히든(hidden) 층을 사용하여야 복잡한 영역(음성 패턴 영역)을 잘 분류할수 있다고 한다. 이번에 사용한 히든(hidden)층의 수는 최대 2개로 정의하였는데 이유는 히든(hidden)이 2개만 되면 어떠한 복잡한 패턴도 분류가 가능하다는 결과에 기인한다. 이 히든(hidden)층의 unit수는 실험상에서 가변적으로 선택 되어진다.

위와같은 다층 구조를 가지는 인공 신경망을 가지고 행하는 모음의 분류는 2개의 단계로 분리가 되는데 첫째는 이 인공 신경망이 모음 패턴을 분류할 수 있도록 적절하게 학습을 시키는 단계인 학습 단계이고 나머지 단계는 학습되어진 신경회로망이 적절하게 학습이 이루어 졌는지를 테스트 데이타를 통하여 알아보는 단계인 테스트 단계이다. 학습은 2장에서 밝힌 것과 같이 GDR(Generalized Delta Rule)을 사용하였으며 테스트 단계는 배치 작업으로 행하여 진다.

3. 학습 단계

구현된 인공 신경망을 이용하여 실제 모음을 분류할 수 있도록 학습을 해보았다. 학습은 크게 2가지로 분류 할 수 있는데 이들은 각기 인공 신경망에서의 입력 층, 출력 층,hidden 층의 unit수를 변경함과 학습 상수를 변경 함으로써 분류하였다.

	에러 한계	학습 비	바이어스 사용	모멘텀 사용	히든층 수	바이어스 학습비
기본형태	0.14	0.4	사 용	사 용	1 층	0.05
변형 1	0.14	0.35	미사용	사 용	1 층	
변형 2	0.14	0.5	미사용	미사용	1 층	
변형 3	0.10	0.45	사 용	미사용	1 층	0.05
변형 4	0.14	0.4	사 용	사 용	2 층	0.05

표 1. 실험 1의 학습 사양

학습 1에서는 입력 층 40개, 출력 층 8개, hidden 층 52개를 가지는 인공 신경망을 사용하여 여러가지 상수(constant)값을 변경하여 비교해 보았다.이때의 상수 값은 에러 바운드(error bound),학습 비(learning rate),바이어스 사용,모멘텀의 사용 등이 된다. 학습 2에서는 제한된 데이타에 대하여 어떻게 hidden 층이 반응하는가를 조사해 보기 위해서 입력 층과, 출력 층의 수를 같게하고, hidden 층의 unit수를 $\log_2$(입력 unit수)로 하고 학습을 시켜 봄으로써 히든 층이 어떻게 입력을 처리 하는지를 간단하게 살펴보고자 한다. 위의 모든 실험은 IBM/PC AT에서 "C" 언어를 사용하여 시뮬레이션을 수행하였다.

1). 학습 1

학습 1에서는 신경회로망 구조를 입력 층을 40개,출력 층의 수를 8개,히든 층을 52개로 고정한 것을 기본으로 하여 인식율을 측정하고 이외에 상수를 변경함으로써 학습의 정도를 분석하였다. 상수의 변경은 5번을 하였는데 그 각각에 대한 사양을 아래에서 표 1로 나타내 보았다.

2). 학습 2

학습 2에서는 제한된 영역에서의 히든 층의 역할을 조사해보기 위한 것이다. 이 학습 3 에서는 입력 층과 출력 층을 같게한다. 따라서 입력 패턴과 출력 패턴은 모양이 동일 하다. 히든 계층은 $\log_2$(입력 계층수)로 한다. 따라서 히든 층의 unit수는 6개가 된다. 이와같은 문제를 엔코딩 문제라 할 수있는데 이유는 입력과 출력의 패턴을 하나의 이진 벡타(binary vector)라 한다면 이들 동일한 벡타를 매핑 시키는 것이기 때문이다. 이러한 환경에서 히든 층이 어떻게 입력 패턴과 출력 패턴을 연결 시키는 지를 알아봄 으로써 간접적으로 히든 층의 제한된 영역에서의 역할을 정의해 보는 것이다. 이 학습에 사용되어진 상수를 표 2로 나타내면 아래와 같다.

4. 실험의 결과 및 분석

학습 단계에서는 각 실험을 통하여 인공 신경망을 패턴을 분류할 수있도록 학습을 시켜 보았다. 이들 학습을 한 인공 신경망이 학습 데이타가 아닌 테스트 데이타도 잘 분류할 수 있는지를 검사하고 이를 결과를 분석해 본다.

1) 학습 1

학습 1에서의 기본형태에 대한 학습 되어진 웨이트를 사용하여 테스트 데이타를 통해서 인식율을 조사해본 결과 아래 표 3과 같이 94%의 인식율을 얻었으며 주된 오류는 /에/와 /애/ 발음의 오류였으며 이는 사람에게서도 나타나는 오류이다.

error bound(ecreit)	0.14
learning rate(eta)	0.4
bias 사용 여부	사용
bias learning rate	0.05
momentum 사용 여부	사용

표 2. 실험 2에서의 학습 형태

실제음 \ 분류음		아	어	오	우	으	이	에	애
아	16	16							
어	16		16						
오	16			15	1				
우	16					15	1		
으	16					1	15		
이	16							16	
에	16							13	3
애	16							2	14
						계 : 117			
						인식율 : 94 %			

표 3. 학습 1의 기본 패턴에서의 인식율

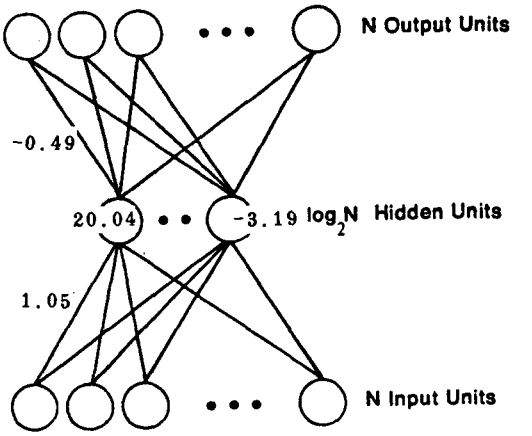


그림 6. 은닉부의 바이어스와 가중치

학습 1에서의 나머지 5가지의 음선에 의한 학습에서 나타난 기본 폴과의 수렴 정도를 그림 5를 통하여 보면 주어진 기본 형태에서 learning cycle의 횟수에 따르는 error값의 변화를 볼 수 있으며 이러한 변화를 통하여 상수의 변화에 따른 학습의 효과에 대해서 비교해 볼 수 있다

2). 학습 2

학습 2에서는 제한된 데이타에서의 히든 층의 역할에 대하여 알아보는 것이다. 이번 실험에서 학습이 되어진 후에 히든 층의 각 unit이 나타내는 바이어스 값과 특정 입력 unit과 출력 unit과의 가중치를 알아보면 그림 6과 같다. 그리고 각각의 입력에 대한 히든 층의 각 unit의 출력값을 테이블로 나타내면 표4와 같은데 이 테이블은 각각의 히든 층 unit의 값이 0.9이상이면 1로 0.1미만이면 0으로 나타내었다. 여기에서 히든 층의 역할이 각 모음의 구분을 위한 적절한 feature를 발견하는 것을 볼 수있다. 즉 히든 unit U_1은 후설 모음및 저모음에 대하여 on이 된다. 그리고 전설 모음에 대해서는 U_3, U_4, U_6이 모두 on이 된다. 이와 같이 히든 층이 인식에 적절한 feature를 자동 추출하는 것이다.

V. 단어 인지 실험

이 실험은 trace model의 심리학적인 언어 인지 현상을 시뮬레이션한 것으로 각 layer간에 interactive하는 동작과정을 IAC model을 통하여 구현하였다. 본 실험의 목표는 7개의 feature들에 대해 정의된 가상의 음성을 입력으로 받아들여 단어 사전에 있는 단어로 수렴되는 과정을 보이는 것이다.

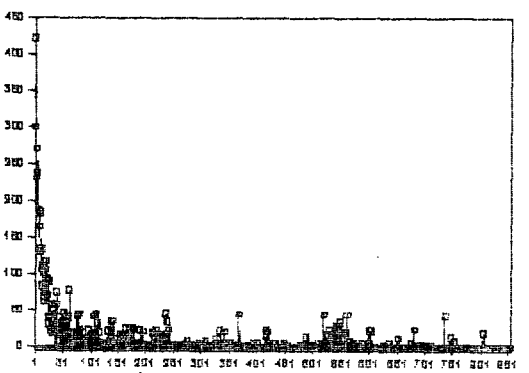


그림 5. Learning Cycle 횟수에 따른 error값 변화

1. IAC(Interactive Activation & Competition) Model

이 모델은 trace model에서 사용하고 있는 interaction개념을 이용하여 구현한 것 으로 McClelland & Rumellhart가 1982년에 제안한 것이다. 그 Network Architecture 는 3개의 processing level-feature, letter, word-로 구성되어 있다. 이러한 level간의 connections은 feature-to-letter excitation, feature-to-letter inhibition, letter-to-word excitation, letter-to-word inhibition, word-to-letter excitation으로 구성되어 있다. 이 모델의 입력은 feature level units이 turn on되었는지 아닌지에 대한 이진 벡터 형태이다. 이러한 입력을 가지고 실제의 모델에서의 처리는 interactive activation and competition mechanism 을 통해서 일어나는데 그 자세한 과정은 다음과 같다.

먼저 두 unit a와 b 각각의 excitatory input을 e_a, e_b라 표시할 때, $e_a > e_b$이고 r은 the strength of the inhibition이라고 가정한다면 각 unit에 대한 net input net_a,net_b은 다음 식으로 표시될 수 있다.

net_a(a에 대한 net input) = e_a - r($output_b$)
net_b(b에 대한 net input) = e_b - r($output_a$)

그런데 각 units의 activation이 양이라면 $output_i = a_i$(단 i는 unit number)이므로 위 식은 다음과 같이 바뀔 수 있다.

$net_a = e_a - ra_b$, $net_b = e_b - ra_a$

따라서, unit b는 inhibition : the rich get richer에 의해서 a가 경쟁에서 이기게 된다. 한편, 일반화 시키면 $net_i = w_{ij}o_i$이므로 net input이 excitatory이면 해당 unit의 activation이 증가하는 방향으로, inhibitory이면 감소하는 방향 으로 작용한다. 즉, 다음과 같은 식으로 표시된다.

If ($net_i > 0$) $a_i = (max - a_i)net_i - decay(a_i - rest)$
otherwise $a_i = (a_i - min)net_i - decay(a_i - rest)$

입력 모음	은닉부 units					
	u1	u2	u3	u4	u5	u6
아	0	1	1	1	1	0
어	0	0	0	1	1	1
오	1	1	1	1	0	0
우	1	0	1	1	0	1
으	1	0	0	0	0	0
이	0	0	1	1	1	1
에	0	0	1	1	0	1
애	0	1	1	1	0	1

표 4. 은닉부의 각 unit의 값

가구[gagu]	가수[gasu]	가시[gasi]
구기[gugi]	구두[gudu]	구리[guri]
기구[gigu]	기사[gisa]	기타[gita]
바다[bada]	바라[bara]	발라[bala]
부두[budu]	부부[bubu]	부리[buri]
비리[biri]	비서[bis^]	비파[bipa]
사수[sasu]	사기[sagi]	사부[sabu]
수구[sugu]	수라[sura]	수리[suri]
시기[sigi]	시사[sisa]	시비[sibi]

표 5. 단어 사전

2. 실험예

본 실험 모델의 입력 데이타로는 각 음소에 대한 특징 벡터를 8개의 level로 구성된 7가지 특징으로 분류하여 모두 56bit의 가상의 이진 벡터로 정의하였다. 예를 들면 음소 [p]의 경우에는 power, vocalicness, diffuseness, acuteness, consonantal, voicing, burst amplitude의 7가지 특징 부분이 각각 4,1,7,2,8,1,8로 정의 되어 2진 값으로 입력되게 된다. 그리고 4개의 음소로 구성된 27개의 단어를 정의하여 표 5와 같이 저장하였다. 이렇게 입력된 음소와 단어들은 계속적인 interaction과 competition에 의해 언어 인지 현상의 일부를 보여주는데 그중의 하나인 lexical effect의 예가 그림 7에 있다. 단어 사전에는 [구두 /gudu]라는 단어가 있는데 [쿠두 /kudu]가 입력되었을 경우에 처음에는 [k] 음소가 많은 score를 얻었으나, 실행회수가 증가함에따라 [g] 음소값이 증가하여 결국 [구두 /gudu]라는 단어로 인식하게 된다. 기타 categorical perception 및 단어 초기 부분의 왜곡의 극복 등의 실험도 행하였다.

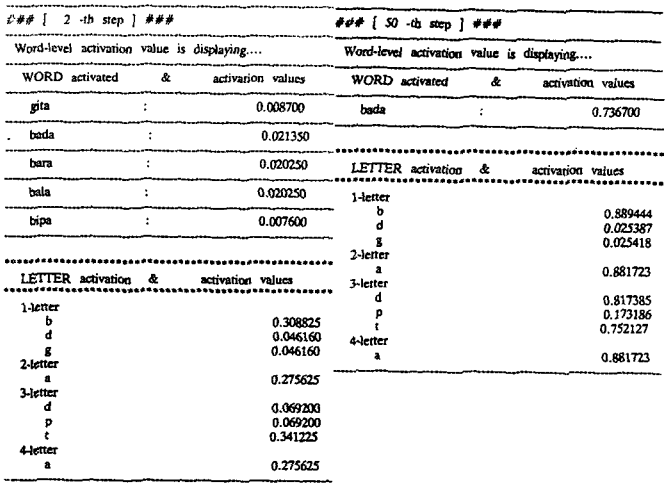


그림 7. Lexical Effect (입력: 바라[bara])

VI. 결론

본 연구에서는 한국어 음성인식시스템의 구현에 인공 신경망 모델을 도입하고자 하는 노력의 일환으로 한국어 단모음의 분류를 Multi-layer Perceptron을 이용하여 행하였다. 1인의 남성 화자가 발음한 제 1,2,3 포르만트의 값을 입력으로 사용하여 약 94%의 인식율을 얻었다.

그리고 또한 상위 레벨의 모듈을 구현하기 위하여 인지 실험을 또 하나의 인공 신경망 모델인 IAC 모델을 통하여 실험하였다. 이 실험을 통하여 한국어 단어 20여개에 대하여 왜곡된 음성 피쳐의 입력에 대한 적응력, lexical Effect, Categorical Perception등의 인지 현상에 대한 실험을 행하였다.

본 연구는 한국어 음성 인식에 인공 신경망 모델을 사용하려는 실험적 연구이며, 본 연구 결과를 바탕으로 한국어 음소 인식 시스템 및 음성 이해 시스템의 구현을 위해 노력할 것이다. 먼저 위 실험을 여성 화자를 포함한 다수의 화자 음성을 대상으로 하여 인공 신경망 모델이 기존의 화자 독립 시스템에서의 Invariant Cue의 추출 노력을 자동 학습으로 대체하는 효과를 뛰어나게 행할 수 있는지를 실험하여야 한다. 그리고 기타 연속 음성상에서의 모음 및 자음의 추출 등에 대한 것도 앞으로의 연구 방향이 될 것이다. 앞으로 궁극적인 목표를 위해서는 이러한 인공 신경망 모델의 연구와 함께 한국어 음성에 대한 음성학적인 연구 그리고 인간의 청각 시스템에 대한 면밀한 연구가 진행되어야 할 것이다.

VII. 참고 문헌

[1] 전상범, 영어 음성학 ,을서 문화사,1983.

[2] 김재하, 음소 분류를 위한 한국어 고립단어 인식에 관한 연구 ,한국과학원 석사 논문,1986.

[3] 이 현복, 한국어의 표준 발음,대한 음성 학회,1987.

[4] 김 기석, "음성이해 시스템에서의 신경회로망 구조의 고찰",정보 과학회지 기술 해설서,1988.

[5] 이 기동, 신 현숙 "언어학 개론", 한신 문화사.

[6] 김 기석, 최 윤석,황 희웅 "한국어 단모음의 분류에 관한 연구",한국정보 과학회 논문지,1987.

[7] D.Rumelhart,G.Hinton,and R.Williams, "Learning internal representations by error back propagation", Parallel Distributed Processing vol.1 pp. 318-362 MIT press, 1986

[8] D.Rumelhart,G.Hinton,and R.Williams, "Learning representation by Back-proragation errors", Nature 323 pp. 533-536, 1986

[9] Masanao Aoki, "Introduction to optimization techniques", Macmillian Company, 1971

[10] Dr Kamal karma and David Breen, "An artificial neural networks tutorial : part 1-basics", Neural Networks, vol 1, No 1, pp. 4-22, january 1989

[11] R.Lippman, "An Introduction to computing with neural nets", IEEE ASSP, pp. 4-22, April 1989

[12] D.R.Rush and J.M.Salas, "Improving the learning rate of Back-propagation with the gradient reuse algorithm", IEEE International conference on neural network, pp.441-447, 1988

[13] J.A.Feldman and D.H.Ballard, "Connectionist Model and Their Properties",Cognitive Science, vol.6,205-254,1982.

[14] M.Minsky, and S.Papert, Perceptrons: An Introduction to Computational Geometry",MIT Press,1969

[15] T.Sejnowski and C.R.Rosenberg,"NETtalk: A Parallel Network That Learns to Read Aloud",Cognitive Science,vol 10,1986.

[16] Jean-Paul Hanton,"Automatic Speech Analysis and Recognition",Preceedings of the NATO Advanced Study Institute,France,June 29 - July 10,1981.

[17] Arbib, "Brains,Machines and Mathematics",1987.

[18] J.L McClelland and J.L.Elman, "Interactive Processes in Speech Perceptron:The Trace Model, Parallel Distributed Processing vol.2 pp. 59-121.

[19] Victor R. Lesser and Lee D Erman,"A retrospective view of the Hearsay II.Architecture", Proceedings 5th Int. Joint Conf. AI,Cambridge Mass.,1977,pp.790-800.

[20] Dennis H. Klatt,"Review of the ARPA Speech Understanding Project",J.Acoust.Soc.Am. vol. 62,No 6,December 1977.