신경망을 이용한 우리말 음성의 인식에 관한 연구
-복합 신경망을 이용한 초성자음 인식에 관한 연구-

A Study on the Word Recognition of Korean Speech using Neural Network
-A Study on the Initial Consonant Recognition using Composite Neural Network-

김 석 동* 이 행 세**

(Sukdong Kim* Haingsei Lee**)

要 約

본 논문은 신경망을 이용한 자음인식에 관한 연구이다. 우선 자음과 모음이 포함된 음성에서 자음부분을 분리하였다. 각각의 자음을 몇개의 집단으로 나누어서 자음구간대 영교차율을 조사하였다. 마지막으로 자음을 인식하기 위해 제어망과 몇개의 소규모 망으로 구성한 혼합 신경망을 제안한다. 제어망은 입력된 자음이 어느 집단에 속하는가를 결정하고, 소규모망에서는 각 집단에 속하는 자음을 인식한다.

ABSTRACT

This paper is a study on the consonant recognition using neural network. First, the part of consonant was separated from the sound of vowel and consonant by the use of acoustic parameters. The rate of length vs. zero crossing rate in the sound of consonant had been studied by dividing each consonant into several groups. Finally, for the purpose of consonant recognition, the composite neural network which consists of a control network and several sub-network is proposed. The control network identifies the group to which the input consonant belongs and the sub-network recognizes the consonant in each group.

이 논문은 1991년도 교육부지원 한국학술진흥재단의 자유공모(지방대학육성)과제 학술연구조성비에 의하여 연구되었음.

I. 서 론

음성은 인간이 사용하는 많은 정보전달 수단중에서 시각과 함께 가장 자연스럽게 이용되고 있고, 사용하는데 특별한 사전지식이나 별도의 훈련이 필요없는 수단이다. 사회의 정보화가 급속히 진전되면서 인간과 기계와의 접촉이 빈번해짐에 따라 인간과 기계사이의 의사전달 방법으로서의 음성의 중요성은 더욱 증대되고 있다. 따라서 음성을 자동으로 인식하고 이해하는 기술을 개발하는 것이 본연구의 궁극적인 목표이다.

또한 음성은 나라마다 음운의 특징이 다르고 사용빈도나 사용음운이 서로 달라서 우리 스스로 극복해야할 요소가 많이 포함되어 있다. 또한 발성자를 한정시키지 않고 연속적으로 발음한 음성을 인식하기 위해서는 인식의 단위를 단어이하의 음소와 같은 비소단위로 해야하는데 이러한 음소단위인식을 위해서는 개인차의 분제와 음소간의 조음결합이 가장 큰문제로 지적되고 있다. 그런데 단모음이나 독립적인 자음을 인식의 대상으로 한정시킨다면 조음결합에 관계없이 개인차만을 고려할 수 있다. 음소단위의 음성인식에 기초가 되는 단모음에 대한 인식을 인공신경회로망의 학습기능을 이용하여 실험한 결과를 발표한바 있으며[1], 초성자음중에서 유성파열음인(/ㄱ/, /ㄷ/, /ㅂ/)를 신경회로망의 구조에 따른 인식률의 변화에 대하여 발표한 바가 있다[2]. 일반적으로

*호서대학교 전자계산학과
**아주대학교 전자공학과
접수일자 : 1992. 3. 4.

자음은 모음과 달리 주파수가 높고 에너지가 낮아서 잡음과 유사한 특성이 있으므로 잡음과 자음, 자음과 모음을 명확히 구분하기가 어렵다. 본 연구는 모음이 뒤따르는 초성자음의 경우에 대하여 자음에서 모음으로 변화되는 천이부분의 저주파 성분을 이용하여 자음과 모음을 분리하였다. 자음을 인식하는 방법으로는 근래에 들어 패턴인식의 도구로 자주 이용되는 신경회로망을 사용하였다. 음성을 인식하기위해 신경망을 적용할때 보통 3가지의 주요한 문제가 대두된다. 첫째로 상당히 많은 음성을 인식하기 위해서는 훈련시간과 훈련 데이타의 갯수를 얼마로 할것인가? 둘째로 신경망이 새로운 어휘를 수용할 수 있게 하기 위한 유연성이 있는가? 마지막으로 시간에 따라 변화하는 음성의 특성을 올바르게 나타낼 수 있는가에 대한 문제가 제기된다. 음성의 특성을 나타내는 여러가지 기술이 발표되어 있다. 시간축과 특징파라미터의 2차원 패턴으로써 음성을 취급하는 방법[3]이 있으며, 특히 초성자음에 대해서 지연회로를 이용하여 시간과 무관한 특징을 추출하는 방법[4]있다. 시간 지연 신경망(Time-delay neural network)을 이용한 초성자음 인식방법은 시간을 이동하면서 인식된 결과를 누적하여 최종적으로 음절을 인식하는 방법이나 본 논문에서는 음성을 연속적인 특징파라미터의 형태로 다루어 우선 자음을 분리하고 분리된 자음을 음소별 인식을 하였다. 또한 오류 역전파 (Error back propagation)에 의한 자음의 인식률과 훈련문제와 신경망의 유연성에 대하여 살펴보겠다. 오류 역전파망은 Rumehalt등이 제안한 방법으로 오차를 거꾸로 전파시키면서 학습규칙을 훈련하는 알고리즘이다. 그 음성인식분야에서 적은수의 어휘를 인식하는데 오류 역전파망은 높은 인식률을 얻을수 있다고 발표되어 있다.[6] 그러나 어휘가 많은 음성을 일반적인 오류 역전파망으로 적용하면 훈련시간이 망의 크기에 비례하여 매우 커지게 된다는 단점이 있다. 더구나 커다란 망내에서는 해(solution)공간내의 총체적인 최소점보다는 부분적인 최소점을 발견할 확률이 높아 최적의 해를 구하기가 매우 어렵다. 또한 새로운 어휘를 추가할때 전체적인 망을 재훈련 시켜야 하는 반복적인 문제가 발생한다. 그러므로 일반적인 오류 역전파망으로는 많은 어휘에 대해서는 실효성이 없게된다. 이러한 문제점을 보완하기 위해서 본 논문에서는 전체 인식대상을 몇개의 그룹으로 나누어 훈련시키는 방법을 사용하였다. 부분망을 사용하므로 해서 그 그룹에 속하는 데이타만을 훈련시키므로서 훈련시간을 감소시킬 수 있으며 새로운 어휘에 대해서도 손쉽게 확장할 수 있는 장점을 가진다. 음성을 컴퓨터에 의해 자동적으로 인식하기 위해서는 음성신호를 분석하여 얻은 파라미터를 인식의 기본 요소로 사용한다. 음성인식을 위한 특정 파라미터 추출 방법에는 크게 시간영역분할방법[1]과 주파수영역 분석방법[2]으로 나눌 수 있는데, 시간영역 분석방법에는 영교차율, 음성에너지, 선형 예측 계수들이 있고 주파수 영역방법에는 포만트(formant)주파수, 켑스트럼(cepstrum)분석, 필터뱅크분석등이 있다. 본 연구는 파라미터 추출방법으로 시간영역분석방법을 이용하였고, 자음을 인식하기 위한 사.모 분리는 자기상관과 에너지및 영교차율을 이용하였다.

Ⅱ. 유음구간 검출

연속적으로 발음된 음성을 단어별로 분할하는것, 즉 음성의 시작점과 끝점을 검출하는 것을 유음구간 검출이라고 한다. 고립단어의 자동 인식에서 단어에 일치하는 음성 신호의 영역을 표시하는 것이 중요하다. 음성 신호의 유음구간 검출은 음성에 해당하는 입력 부분만 처리하도록 함으로써 실 시간이 아닌 시스템에서 계산량을 줄이는데 중요한 역할을 한다. 정확한 유음구간 검출은 S/N비가 높은 환경에서는 가장 낮은 레벨의 음성(에너지가 낮은 마찰음)에너지 조차도 배경 잡음 에너지를 초과한다. 그래서 단지 에너지만으로도 유음구간을 쉽게 검출할 수 있다. 그러나 대부분의 실제적인 음성 시스템은 15~20 dB 정도의 낮은 S/N 비에서 동작된다. 일반적으로 다음과 같은 환경에서는 음성의 시작점과 끝점을 분리해 내기가 어렵다.

1. 음성의 시작과 끝에 에너지가 낮은 마찰음 (/f/, /th/, /h/)이 존재할때
2. 음성의 시작과 끝에 에너지가 낮은 파열음 (/p/, /t/, /k/)이 존재할때
3. 음성의 끝에 비음이 있을때
4. 유성 마찰음이 단어의 끝에서 무성 마찰음으로 될때
5. 단어의 끝에서 모음의 에너지가 낮을때

위와같은 어려움에도 불구하고 에너지와 영교차율은 음성의 유음구간을 검출하는데 유용하다. 본 논문에서는 프레임내의 평균에너지와 피크 신호 크기를 이용해 음성의 유음구간을 검출 하였다. 음성 신호의 통계적 특성이 시간에 따라 거의 변하지 않는

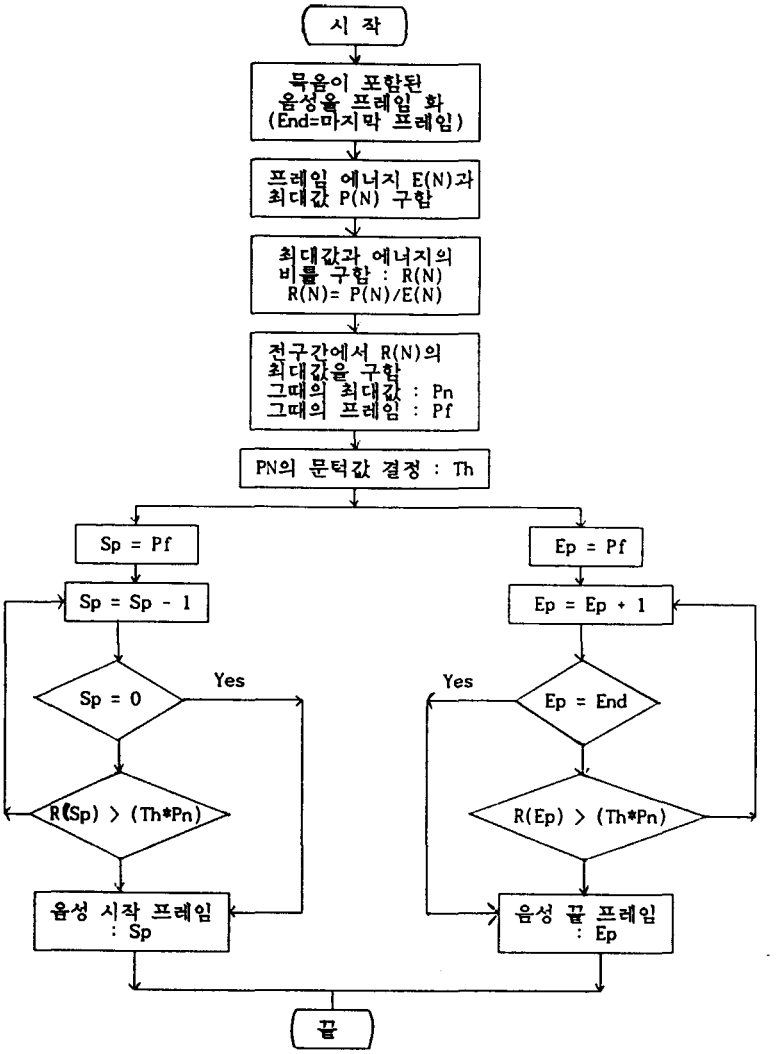


그림 1. 유음 구간 검출 알고리즘

준 정상적(quasi-stationary)인 사실을 이용해 묵음이 포함된 전 음성 구간을 일정한 시간 간격으로 분할(segmentation)하여 프레임(frame)으로 나눈 다음 각 프레임에서 평균 에너지와 피크 신호 크기의 비를 구한다. 두 에너지의 비(rate)가 최대가 되는 프레임을 찾아 양쪽으로 가면서 문턱값(threshold) 이하가 되는 프레임을 음성의 시작점과 끝점으로 한다. 음성의 유음구간을 검출하는 알고리즘을 그림 1에 나타내었다.

Ⅲ. 자음 추출

본논문에서 제안하는 자음부분 추출방법은 크게 두부분으로 구성된다. 첫번째로는 단구간 자기상관(short-time autocorrelation)을 이용하여 잡음성이 강한 자음(무성자음)이 포함된 음성과 모음성이 강한 자음(유성자음)이 포함된 음성을 구분하고, 두번째로는 무성자음부분은 단구간 영교차율을 이용하고 유성자음부분은 단구간 에너지를 이용하여 자음부분과 모음부분을 분리하였다.

1)무성자음 음성과 유성자음 음성의 분류

음성신호를 x(m)이라 할때 단구간 자기상관 함수 $\phi_l(k)$는

$$\phi_l(k)=\sum_{m=-\infty}^{\infty} x(m)\,W(l-m)\,x(m+k)\,W(l-m-k)$$

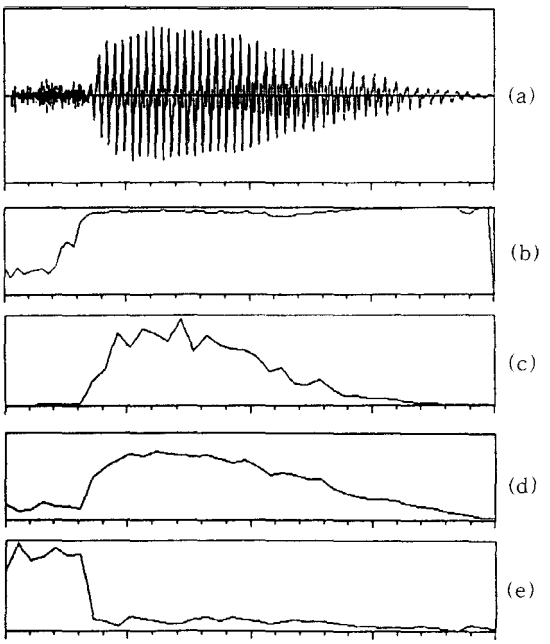


그림 2-1. 음성 /기/의 (a)음성파형 (b)rm (c)단구간 에너지 (d)포락선 (e)영교차율

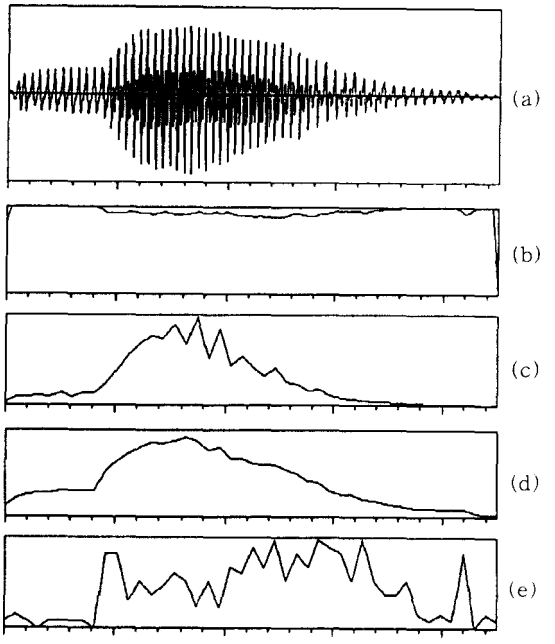


그림 2 2. 음성 /니/의 (a)음성파형 (b)r (c)단구간 에너지 (d)포락선 (e)영교차율

가 된다. 여기서 l은 프레임을 나타내고 k는 지연시간이고 N은 프레임의 길이이며 W는 창(window)함수이다. 본논문에서 사용한 창함수는 해밍창으로

$$W(n)=\begin{cases}0.54-0.46\,COS(2\pi n/(N-1)), & 0\le n\le N-1\\ 0, & \text{otherwise}\end{cases}$$

이다. 1번째 프레임에서의 $\phi(0)$를 제외한 최대 자기상관값을 r이라 하면

$$r=\underset{k=1..N-1}{\text{argmax}}\{\phi(k)\}$$

이다. 여기서 argmax{}는 괄호내의 값중 최대가 되는 것을 선택하는 함수이다. 일반적으로 유성음과 무성음에 따라 r값이 다르다. 그림 2-1은 무성자음이 포함된 음성 /기/에 대한 음성이며, 그림 2-2는 유성자음이 포함된 음성 /니/의 음성파형, r, 에너지, 포락선과 영교차율을 보이고 있다. 그림에서 볼 수 있듯이 에너지와 영교차율은 작고 r이 클때는 유성자음이고, 에너지와 r은 작고 영교차율이 클때는 무성자음임을 알 수 있다.

2)자음과 모음의 분류

우리말의 자음은 혼자서는 스스로의 음가를 발휘하지 못하고 모음과 합쳐서 발음해야 한다. 본 논문에서는 자음부분만을 인식하기 위해서 모음과 같이 발음한 음성중에서 자음부분만을 분리해 내야 한다. 음성에서 자음과 모음을 정확하게 분리하는 것은 음성인식처리 과정에서 해결해야 할 중요한 일이다. 본 논문에서 사용한 자.모음의 결정방법은 일종의 패턴인식방법으로 자음과 모음사이의 음성학적 특징을 이용하였다. 첫째로 무성자음과 모음의 분리방법은 다음과 같다. 파열음이나 마찰음과 같은 무성자음은 잡음과 비슷하여 주파수가 높으면서 에너지가 낮다. 이를 이용하여 본 논문에서는 자.모 분리를 우선 고주파 성분의 백색잡음이나 팝 노이즈와 같은 저주파 성분을 제거하기 위해 음성신호를 band-pass여파기를 통과시키고나서, 자음에서 모음으로 변화되는 천이구간에서의 작은 저주파 성분을 이용하여 자.모음의 분리를 하였다.

둘째로 유성자음과 모음의 분리방법은 다음과 같다. 비음과 같은 유성자음은 모음과 비슷한 특징을 가지며 에너지와 주파수가 약간 작다. 음성 전구간에 대하여 단구간에너지를 구하면 모음부분에서 최대값

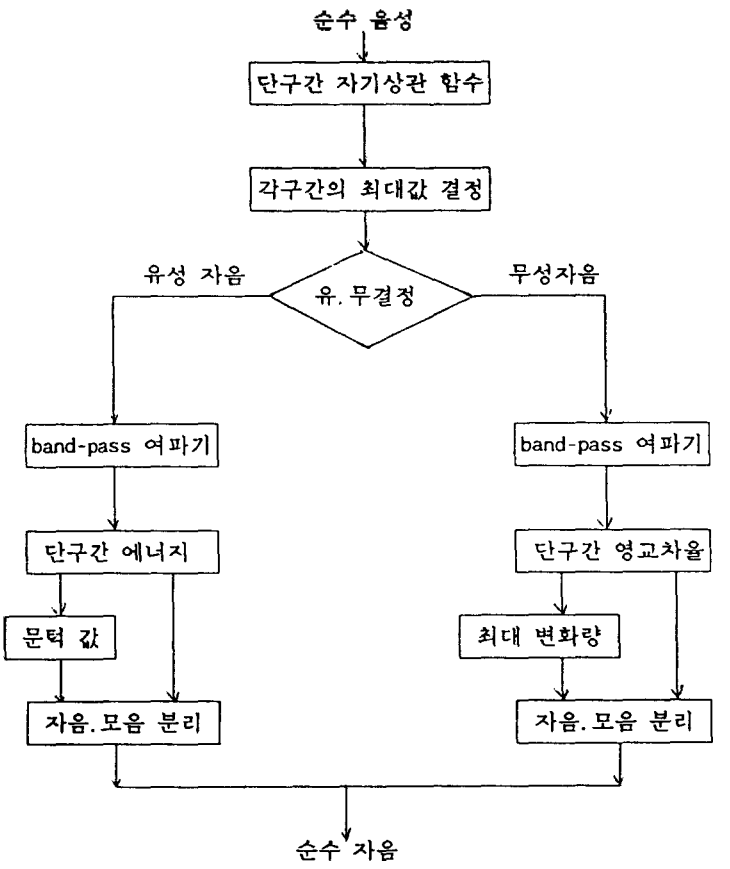


그림 3. 자음 부분 추출 방법

을 갖는 프레임이 나타난다. 최대값이 나타나는 프레임에서 시작하여 앞 프레임으로 이동하면서 문턱값(Threshold value)보다 작은 값을 갖는 프레임이 자음부분임을 알 수 있었다. 그림 2-1에서 보는것 같이 에너지가 매우 작은 앞부분이 자음임을 알 수 있다. 그림3에 자음 부분 추출 방법에 대한 전반적인 알고리즘을 나타내었다.

3)클러스터링

본 논문에서는 인식율을 높이기 위해 유사한 특성을 갖고 있는 음성들끼리 음성을 분류했다. 음성을 분류하는 방법에는 여러가지 이론이 있는데 조음의 방법에 따라 경음(pressure) 비음(nasal sound) 파열음(plosive sound), 마찰음(fricative sound) 등으로 분류하고, 조음 위치에 따라 인두음(larynx sound) 연 구개음(velar sound) 구개음(palatal sound) 등으로 분류한다. 그러나 본 논문에서는 각 음성별 특질을 사용해서 음성을 분류했다. 음성을 분류 하는데 전 음성중에서 자음 구간까지 프레임별 구한 영 교차율의 평균치와 자음 구간의 길이 분포를 특질로 사용해 음성을 분류했다. 각 음성별 자음 구간까지의 영 교차율 평균값과 자음 구간의 길이 분포를 그림 4에 나타내었다. 자음구간은 프레임의 수를 나타낸것으로 프레임은 10 msec의 크기이며 8 msec가 중복이 되도록 하였다. 이 그림에 나타난 분포를 살펴보면 우선 유성자음과 무성자음의 차이가 두드러진다. 유성자음은 그림 4-2에서와 같이 영교차율의 평균값이 작아 그림의 왼쪽에 분포하고 있다. 그러나 무성자음은 모두 오른쪽에 분포함을 알 수 있다. /ㄱ/,/ㄷ/,/ㅂ/는 자음구간이 비교적 짧아 아래 영역에 분포하며, /ㅅ/,/ㅈ/,/ㅊ/는 자음구간이 길므로 가장 높은 영역에 분포하며, /ㅍ/,/ㅌ/,/ㅋ/,/ㅎ/는 중간 영역에 분포함을 알 수 있다.

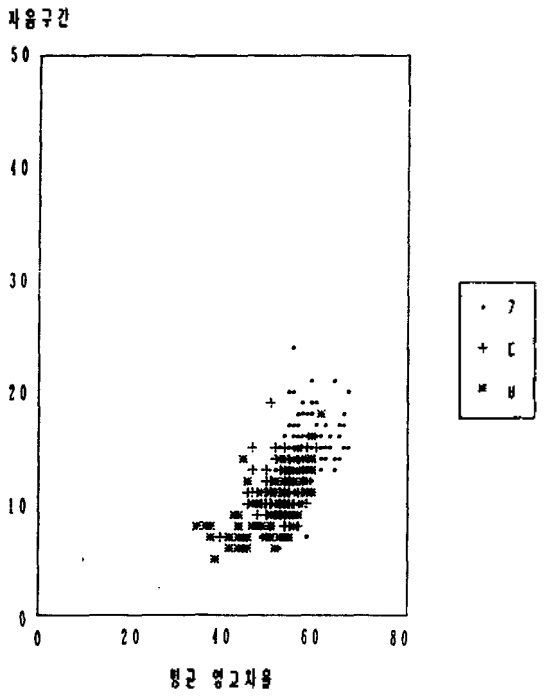


그림 4-1. 자음(ㄱ, ㄷ, ㅂ)에 대한 자음구간과 영교차율의 분포

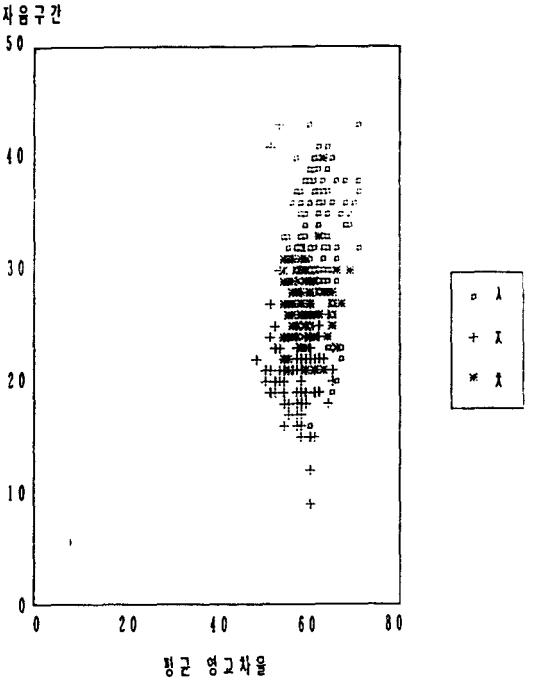


그림 4-3. 자음(ㅅ, ㅈ, ㅊ)에 대한 자음구간과 영교차율의 분포

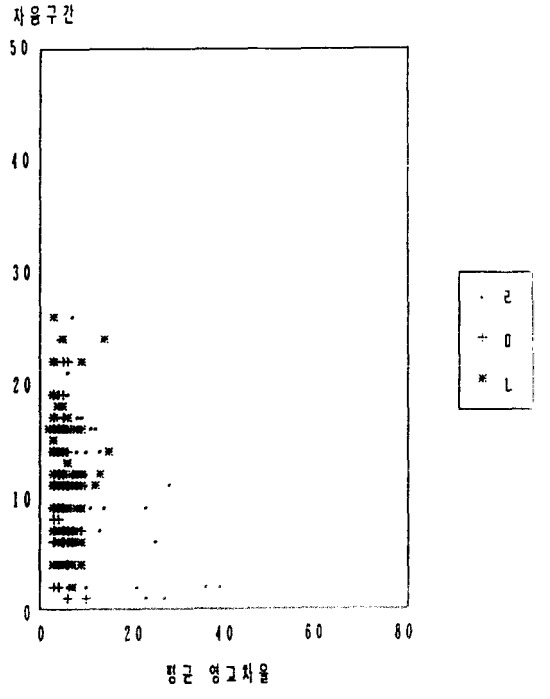


그림 4-2. 자음(ㅋ, ㅌ, ㅍ)에 대한 자음구간과 영교차율의 분포

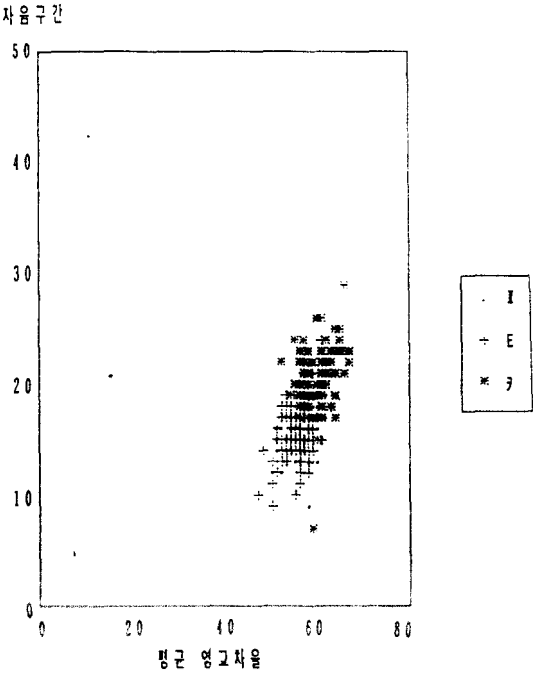


그림 4-4. 자음(ㅎ, ㅁ, ㄴ, ㅇ, ㄹ)에 대한 자음구간과 영교차율의 분포

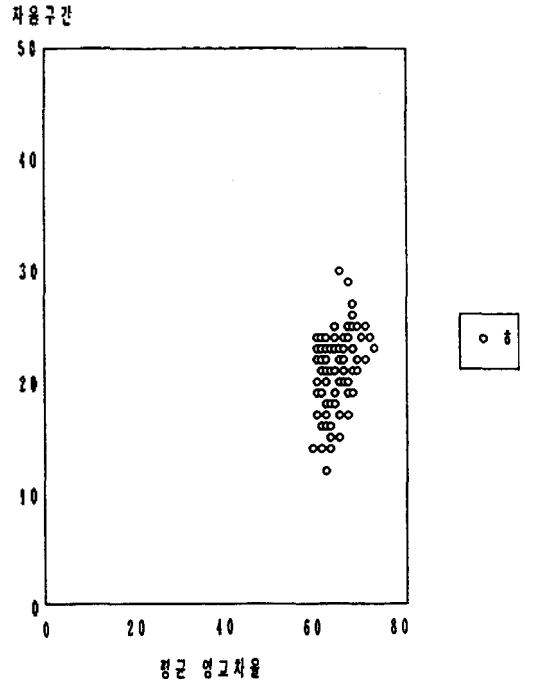


그림 4-5. 자음{/ㅎ/}에 대한 자음구간과 영교차율의 분포

Ⅳ. 자음 인식

연속적으로 발음한 음성을 인식하기 위해서는 인식의 단위를 단어이하의 음소와 같은 미소단위로 해야하는데 이러한 음소단위인식을 위해서는 개인차의 문제와 음소간의 조음결합이 가장 큰문제로 지적되고 있다. 그런데 단모음이나 독립적인 자음을 인식의 대상으로 한정시킨다면 조음결합에 관계없이 개인차만을 고려할 수 있다. 일반적으로 자음은 모음과 달리 주파수가 높고 에너지가 낮아서 잡음과 유사한 특성이 있으므로 잡음과 자음, 자음과 모음을 명확히 구분하기가 어렵다. 본연구는 모음이 뒤따르는 초성자음의 경우에 대하여 자음에서 모음으로 변화되는 천이부분의 저주파 성분과 에너지를 이용하여 자음과 모음을 분리하고, 음성을 연속적인 특징파라미터의 형태로 다루어 음소별 인식을 하였고, Back-propagation에 의한 자음의 인식률과 훈련방법문제에 대하여 살펴보겠다. 전반적인 인식방법을 그림 5에 나타내었다.

1)학습 방법

신경망에 의해 패턴을 인식할 때의 커다란 문제점 중의 하나는 훈련시간을 가급적 적게 하면서 인식률을 높이는 것이다. 훈련시간을 줄이는 방법으로는 여러가지가 있다. 총체적인 오차가 제일 적은 곳 즉 최적해를 찾아가는 수렴 속도를 증가시키는 방법으로 모멘텀을 사용하거나, 훈련할 데이타를 선별하거나, 인식할 패턴에 가장 알맞는 신경망 모델을 선택하는 방법등이 있을 수 있다. 본 논문에서는 훈련시간을 결정하는 훈련횟수를 줄이기 위해 우선 비슷한 특성을 갖는 데이타를 묶어서 여러개의 부분집합으로 나누고, 각 부분집합에 속한 데이타만을 대상으로 훈련하는 방법을 사용한다. 신경망내의 각 층을 연결하는 가중치(weight)의 갯수는 [입력층의 노드수 * 은폐층의 노드수 * 출력층의 노드수]이므로 모든 데이타를 한꺼번에 훈련할때 은폐층의 노드수를 10개라면 1820개이고, 부분적으로 훈련할때의 갯수는 1190개로 65%정도 감소한다. 훈련횟수를 [가중치 갯수 * 훈련 데이타 갯수 * 훈련 반복 횟수]라면 표 1에 1에서 볼 수 있듯이 부분적인 훈련횟수가 전체적인 훈련횟수의 32%로 감소한다.

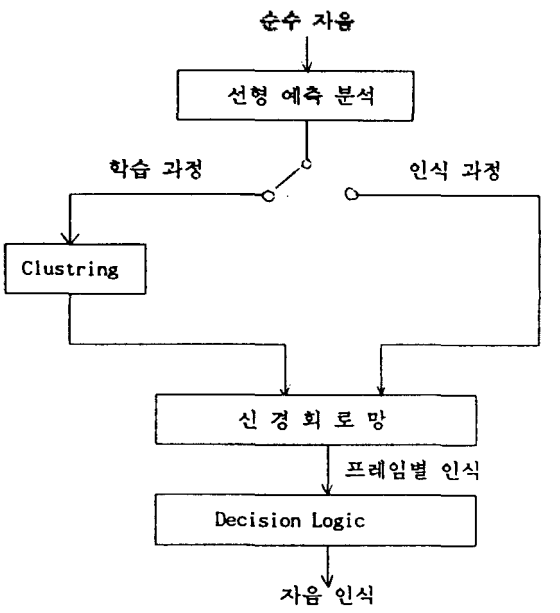


그림 5. 자음 인식에 대한 전반적인 방법

그림 6에 본 논문에 사용한, 규모가 비교적 작은 보통의 신경망이 5개가 결합된 형태의 구조를 보인다. 가장 우측에 있는 신경망(그림에서 B)은 음성이 어느 그룹에 속하는지를 결정하는데 사용하고, 나머지 신경망(그림에서 A1－A4)은 각기 그룹에 속하는 자음을 인식하는데 사용한다. 각각의 부분 신경망은 각각 개별적으로 학습한다. 그룹을 결정하는 신경망은

표 1. 부분적인 훈련횟수와 전체적인 훈련횟수에 대한 비교

	부분적인 훈련방법					전체적인 훈련방법
	ㄱ,ㄷ,ㅂ	ㄴ,ㄹ,ㅁ	ㅅ,ㅈ,ㅊ	ㅋ,ㅌ,ㅍ	그룹결정, ㅎ	
입력층의 노드수	14	14	14	14	14	14
은폐층의 노드수	5	5	5	5	5	10
출력층의 노드수	3	3	3	3	5	13
가중치의 갯 수	210	210	210	210	350	1820
	1190					
훈 련 데이타수	150	150	150	150	650	600
훈련반복 횟 수	2000	2000	2000	2000	2000	2000
총 훈 련 횟 수	$63*10^6$	$63*10^6$	$63*10^6$	$63*10^6$	$155*10^6$	$2184*10^6$
	$707*10^6$					

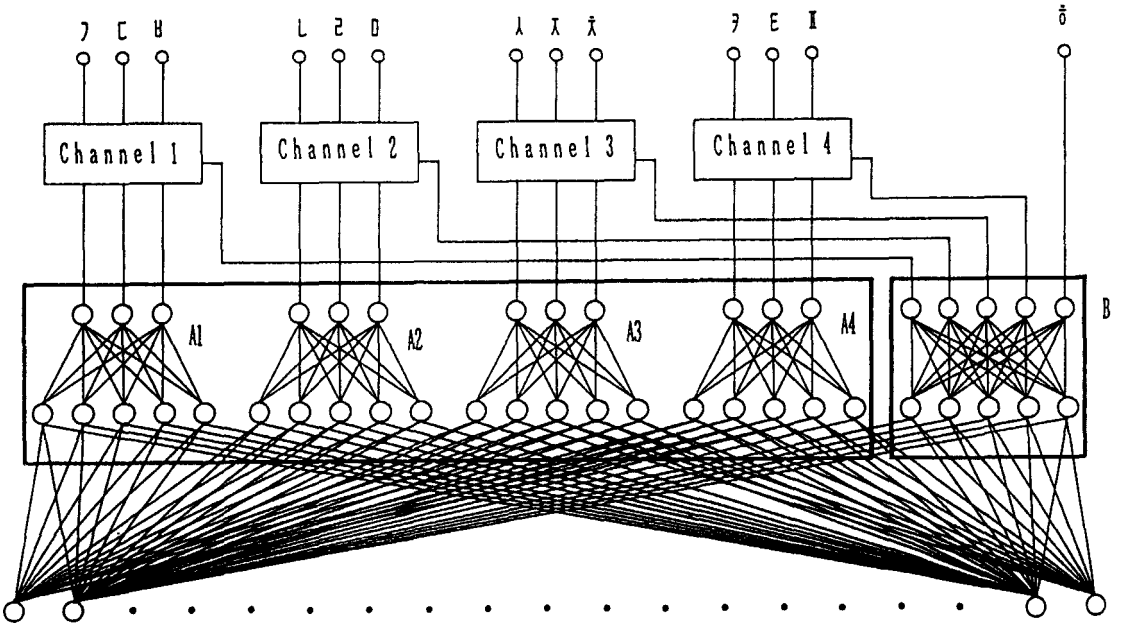


그림 6. 신경망의 구조

모든 자음을 대상으로 학습하고, 자음을 인식하는 부분신경망은 그 그룹에 속하는 자음만을 대상으로 학습한다. 그러므로 각 부분망의 학습시간은 짧아진다. 최종인식결과는 그룹을 결정하는 신경망의 출력과 그룹별 자음을 인식하는 신경망의 출력에 의해 나타난다. 대상 자음은 5개 그룹으로 나누었으나 그룹에 속하는 자음이 하나밖에 없는 /히/는 그룹별 자음을 인식하는 신경망이 필요하며, 왜냐하면 /히/는 그룹을 결정하는 신경망의 출력이 바로 /히/의 인식을 의미하기 때문이다.

다층 신경망의 노드들 사이의 연결 세수를 조정하는 신경망의 알고리즘으로는 실제의 출력 값이 원하는 출력 값을 갖도록 파라미터를 적응 수정하는 알고리즘으로써 D.E Rumelhart, G.E Hinton, R.J. Williams의 BP(back propagation)방법을 이용하였다. 각각의 입력 자료에 대하여 원하는 출력값과 실제 출력값의 차이를 계산하고 그 값은 입력에 궤환시켜 출력오차를 최소하으로 줄이기 위해 모든 노드들 사이의 가중치를 변경한다. 이 모델은 신경세포의 연결에 의한 신호 종합 작용을 나타내며, 노드는 뉴런을 나타내고 연결선은 결합강도를 나타내는 액슨과 시냅스의 성질을 나타낸다. 이를 이용하여 전단 신호가 다음단에서 합해지도록 하였으며 또한 출력이 포화되는 특징을 시그모이드(sigmoid) 함수로 대치한

것이다. 모든 신호는 전향적으로 전파되는 것을 가정하였고 각 연결선의 결합도를 학습에 의하여 적응 하도록 하였다.

2)결과분석

음성 데이타 수집은 DAT(Digital audio tape recoder)로 방음실에서 다이나믹 마이크를 통해서 약 1.6sec 동안 발음한 음성을 자기 테이프에 저장하였다. 사용한 음성은 우리말 자음 가운데 잡음 특성이 강한 유성 파열음(/ㄱ/, /ㄷ/, /ㅂ/), 무성 파열음(/ㅍ/, /ㅌ/, /ㅋ/), 마찰음(/ㅅ, /ㅈ/)과 파찰음(/ㅊ/), 비음(/ㄴ/, /ㄹ/, /ㅁ/) 그리고 /ㅎ/ 등 13개의 자음소와 모음/ㅣ/와의 결합 형태의 음성 데이타를 사용 하였다. 3명의 성인 남성 화자가 각각 30회씩 발음하여 10×3×30=900개의 발음을 수집 하였다. 자기 테이프 재생기와 증폭기를 거쳐 나온 음성을 Data Acquisition Board인 Data Traslation 2801 Board로 음성 크기는 최대치가 ±10V 이내, 샘플링 주파수 10kHz, 양자화 준위 12 bit로 아나로그 디지탈 변환(ADC, analog digital conversion)하여 컴퓨터 기억 소자에 저장 하였다. 신경 회로망의 입력 노드 수는 12차의 PARCOR 계수와 각 음성 구간의 영 교차율 평균치 그리고 자음 구간의 길이를 정규화하여 합한 14개이다. 학습 데이타는 첫번째 사람이 5번 발음한 음성을 대상으로 각각의 자음에 대하여 50개의 프레임을 임의로 선정하여 총 650개의 프레임을 사용하였다. 프레임의 길이는 10ms로 하였고, 자음 구간의 분석 구간을 늘리기 위해 6ms를 중첩 시켰다. 중간 노드의 수는 여러번의 반복적인 실험에 의해 10개로 정하였고, 출력 노드의 수는 그룹을 결정하는 신경망은 5개, 자음을 인식하는 신경망은 3개이다. 학습 횟수는 시간을 제한하기 위해 2,000회 이내에서 각 음성에 대해 실제의 출력값과 원하는 출력값의 차 즉 허용 오차가 0.3 이내에 들도록 하였다. 인식은 두가지 방법으로 조사하였다. 첫째로 모든 프레임에서 인식률을 계산하는 프레임별 인식과, 둘째로 하나의 음성에 대하여 70% 이상 같은 결과가 나올때만 인식률을 계산하는 전 프레임 별 인식을 하였다. 본 실험에 사용한 음성 데이타 내용과 특징 파라미터를 표 2에 제시하였다.

분류율과 자음인식률을 나누어 실험을 하였다. 첫째로 표3에 나타난 것같이 분류율 실험에서 유성 파열음 계통과 마찰음 및 파찰음 계열은 비교적 잘 분류 되었으나 무성 파열음에서는 상호 간섭 현상이 나

표 2. 음성 데이타 채집과 특징 파라미터

대상 자음	/ㄱ/, /ㄴ/, /ㄷ/, /ㄹ/, /ㅁ/, /ㅂ/, /ㅅ/, /ㅈ/, /ㅊ/, /ㅋ/, /ㅌ/, /ㅍ/, /ㅎ/
음성 채집	3명의 성인 남자가 각각 30번씩 발음
샘플링 특성	샘플링 주파수 : 10kHz, 양자화 비트수 : 12bit
분석 구간	프레임 길이 : 10ms, 프레임 중첩 : 6ms, 해밍창 사용
특 정 파라미터	부분 자기 상관 계수(12차), 영 교차율의 평균치, 자음 구간의 길이

표 3. 신경망을 이용한 음성 분류율

음 성	각 프레임 단위별 분류율(%)	전 프레임별 분류율(%)
ㄱ	92.8	100.0
ㄷ	90.5	100.0
ㅂ	90.8	100.0
평 균	91.4	100.0
ㄴ	95.5	100.0
ㄹ	93.5	100.0
ㅁ	96.8	100.0
평 균	95.3	100.0
ㅅ	84.7	100.0
ㅈ	75.6	87.8
ㅊ	82.8	100.0
평 균	84.4	95.9
ㅋ	85.7	100.0
ㅌ	61.8	85.7
ㅍ	60.3	88.2
평 균	69.3	91.3
ㅎ	74.8	90.0
전체적인 분류율	84.3	96.3

표 4. 각 그룹별 인식 결과

1)유성 파열음의 인식율

화자	각 프레임별 인식			화자	전 프레임을 사용한 인식		
	ㄱ	ㄷ	ㅂ		ㄱ	ㄷ	ㅂ
A	87.0	83.7	83.4	A	96.7	96.7	73.3
B	81.9	80.4	70.6	B	96.7	66.7	76.7
C	86.1	74.0	83.7	C	83.3	73.3	80.0
평균	85.1	79.4	79.2	평균	92.2	78.9	76.7
전체 평균		81.2		전체 평균		82.6	

2) 비음의 인식율

화자	각 프레임별 인식			화자	전 프레임을 사용한 인식		
	ㄴ	ㄹ	ㅁ		ㄴ	ㄹ	ㅁ
A	65.9	78.3	93.4	A	72.3	100.0	100.0
B	66.7	76.8	92.7	B	78.8	99.7	100.0
C	70.3	82.5	92.3	C	80.3	100.0	100.0
평균	67.6	79.2	92.8	평균	77.1	99.9	100.0
전체 평균		79.9		전체 평균		92.3	

3) 마찰음과 파찰음의 인식율

화자	각 프레임별 인식			화자	전 프레임을 사용한 인식		
	ㅅ	ㅈ	ㅊ		ㅅ	ㅈ	ㅊ
A	87.6	82.0	90.9	A	90.0	96.7	100.0
B	89.1	89.9	94.3	B	83.3	63.3	90.0
C	97.1	88.6	92.2	C	96.7	70.0	100.0
평균	91.2	86.8	92.5	평균	90.0	76.7	96.7
전체 평균		90.2		전체 평균		87.8	

4) 무성 파열음의 인식율

화자	각 프레임별 인식			화자	전 프레임을 사용한 인식		
	ㅋ	ㅌ	ㅍ		ㅋ	ㅌ	ㅍ
A	84.5	72.9	76.8	A	96.7	93.3	93.3
B	86.6	79.2	75.6	B	93.3	93.3	93.3
C	85.8	80.4	73.1	C	100.0	90.0	80.0
평균	85.6	77.5	75.2	평균	96.7	92.2	88.9
전체 평균		72.6		전체 평균		92.6	

5) /ㅎ/음의 인식율

화자	각 프레임별 인식	화자	전 프레임을 사용한 인식
	ㅎ		ㅎ
A	80.7	A	91.5
B	72.6	B	91.3
C	71.2	C	87.2
평균	74.8	평균	90.0

타나서 분류율이 저조함을 볼 수 있다. ㄱ, ㄷ, ㅂ를 입력으로 해서 분류했을때 프레임 단위 분류에서는 91.1%, 전 프레임을 사용한 분류에서는 100% 분류율을 얻었으며, ㄴ, ㄹ, ㅁ를 입력으로 했을때 프레임 단위 분류에서는 95.3%, 전 프레임 분류에서는 100%, ㅅ, ㅈ, ㅊ를 입력으로 했을때 프레임 단위 분류에서는 84.4%, 전 프레임 분류에서는 95.9%, ㅋ, ㅌ, ㅍ을 입력으로 했을때 프레임 단위에서는 69.3%, 전 프레임 단위로 분류했을때 91.3%, ㅎ을 입력으로 했을때 프레임 단위 분류에서는 74.8%, 전 프레임 분류에서는 90.0%의 분류율을 얻었다.

표 5. 전(全) 프레임별 자음인식률

음 성	부분적인 훈련(%)	전체적인 훈련(%)
ㄱ	92.2	92.5
ㄴ	77.1	70.6
ㄷ	78.9	71.6
ㄹ	99.9	100.0
ㅁ	100.0	95.1
ㅂ	76.7	76.5
ㅅ	90.0	90.1
ㅈ	76.7	77.8
ㅊ	96.7	98.1
ㅋ	96.7	95.1
ㅌ	92.2	91.3
ㅍ	88.9	88.9
ㅎ	90.0	89.7
표준 인식률	88.9	87.5

다음으로 13개의 자음을 3-4절에서와 같이 유사군으로 분류한 후 각 그룹내에서의 음성을 각 프레임 별 인식과 전 프레임을 사용한 인식 실험을 했다. 그 결과가 표 4에 있다. 표 4에 나타난 것처럼 유성 파열음, 비음, 마찰음과 파찰음, 무성파열음, 비음 그리고 /ㅎ/음의 각 프레임 별 인식율은 81.2%, 79.9%, 90.2%, 79.4%, 74.8% 그리고 전 프레임 인식율은 각각 82.6%, 92.3%, 87.8%, 92.6%, 90.0%의 비교적 좋은 인식 결과를 얻었다. 마지막으로 부분적으로 훈련할 때와 전체적으로 훈련했을때의 인식률은 크게 차이 나지 않지만 약간 인식률이 높음을 알 수 있다.

V. 결 론

본 연구에서는 초성자음을 대상으로 음성학적 특성에 따라 분류하고, 각 그룹별 분류율과 인식률을 살펴보았다. 음성별 분류에 해당하는 평균분류율은 96.3%이고, 음소별 분류에 해당하는 평균분류율은 94.3%이며, 전 프레임을 사용한 평균 인식률은 88.1%이고, 음소별 전체 평균 인식률은 82.0%을 얻었다.

인식에 사용한 신경망의 구조를 자음 중의 입력을 대상으로 하기위해서, 또한 새로운 어휘가 나타났을

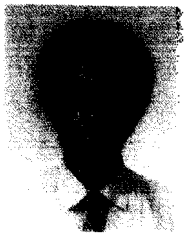
때 훈련시 유연성을 주기 위해 혼합 신경회로망을 제안하였다.

앞으로 전 자음을 대상으로 하기 위해서는 종성 자음까지 확대한 실험이 필요하며 이를 토대로 음소별 인식을 통한 연속음 인식을 실현하기를 기대한다.

참 고 문 헌

1. 김석동, 이행세, "인공 신경회로망의 학습기능을 이용한 우리말 모음인식에 관하여," 신호처리 합동 학술대회 논문집, pp.192-195, 1989.9.
2. 김석동, 이행세, "신경망에 의한 초성자음(ㄱ,ㄷ,ㅂ)의 인식방법," 한국음향학회 학술발표회 논문집, pp.73-77, 1991. 11.
3. T.K.Landaur, C.A.Kamm and S.Singhal : "Teaching a Minimally Structured Back-Propagation Network to Recognize Speech Sounds," Proc. the 9th Annual Conference of the Cognitive Science Society. pp.531-536(1986).
4. A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.Lang, "Phoneme Recognition using Time-Delay Neural Network" IEEE Trans. Vol. ASSP-37, No.8, Aug. 1989.
5. R.P. Lippman, "An Introduction to Computing with Neural Nets," IEEE ASSP Magazine, Vol.4, No.2, pp.4-22, April 1987.
6. D.J.Burr : "Experiments on Neural Net Recognition of Spoken and Writen Text," IEEE Trans. Acoust., Speech, Signal Processing, vol.36, No.7, pp. 11622-1168, July 1988.
7. G.M. White and R.B. Neely, "Speech Recognition Experiment with Linear Prediction, Bandpass Filtering, and Dynamic Programming," IEEE Trans. Acoust., Speech and signal Processing, Vol. ASSP, April 1976, 183-188.
8. H.Hermansky, "an efficient Speaker-Indendent Automatic Speech Recognition by Simulation of some properities of Human auditory perception," ICASSP-87, pp.1159-1162, 1987.
9. H. Kobatake, "Optimization of Voiced/Unvoiced Decisions in Nonstationary Noise Environments," IEEE Trans. Vol. ASSP-35, No.1, January 1987.
10. A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.Lang : "Phoneme Recognition : Neural Networks vs. Hidden Models," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, New York, NY, April 1988.

▲김 석 동(金錫東) 1957년 8월 23일생



1982년 2월 : 아주대학교 전자공학과 졸업(공학사)
1984년 2월 : 아주대학교 대학원 전자공학과 졸업 (공학석사)
1986년 8월－현재 : 아주대학교 대학원 전자공학과 박사과정
1987년 2월－현재 : 호서대학교 전자계산학과 조교수

▲이 행 세(李幸世) 1943년 8월 29일생

1966년 2월 : 전북대학교 전기공학과 졸업(공학사)
1972년 2월 : 서울대학교 대학원 전자공학과 졸업 (공학석사)
1984년 8월 : 고려대학교 대학원 전자공학과 졸업 (공학박사)
1973년 2월－현재 : 아주대학교 전자공학과 교수