주파수 상태 신경 회로망을 이용한 음소 인식

Phoneme Recognition Using Frequency State Neural Network

李 俊 模*, 黃 英 洙*, 金 成 鍾**, 申 仁 澈**
(Joon Mo Lee*, Young Soo Hwang*, Seong Jong Kim**, In Chul Shin**)

요 약

본 논문에서는 음소의 시간 구조 특성만을 다룬 일반적인 TSNN 방법에 음소의 주파수 대역 구조를 포함시킨 신경 회로망을 제안한다. 제안된 신경 회로망에 음소(아, 이, 오, ㅅ, ㅊ, ㅍ, ㄱ, ㅇ, ㄹ, ㅁ)을 학습시켜 인식을 수행한 결과, 시간 인자 특성을 입력으로 음소를 인식한 일반적인 TDNN 방법과 TSNN 방법보다 본 논문에서 시간과 주파수 인자를 동시에 입력으로 수행한 신경회로망 방법이 약간 더 나은 인식 결과를 보였다.

ABSTRACT

This paper reports a new structure for phoneme recognition neural network. The proposed neural network is able to deal with the structure of the frequency bands as well as the temporal structure of phonemic features which used in the conventional TSNN.

We trained this neural network using the phonetics (아, 이, 오, ㅅ, ㅊ, ㅍ, ㄱ, ㅇ, ㄹ, ㅁ) and the phoneme recognition of this neural network was a little better than those of conventional TDNN and TSNN using only temporal structure of phonemic features.

I. 서 론

최근 컴퓨터 기술의 급격한 발전은 상당한 양의 데이터 처리를 가능하게 하고 있으며, DTW(Dynamic Time Warping)[1], VQ(Vector Quantization)[2][3], HMM (Hidden Markov Model)[4], 신경 회로망 [5] 등의 음성 인식 연구 분야에 많은 도움을 주고 있다.

*嶺東大學校 電子工學科
**檀國大學校 電子工學科
접수일자 : 1994년 3월 8일

학습 알고리듬이 개발된 HMM과 신경 회로망은 음소 인식에 널리 이용되고 있으며, 특히 HMM은 연속 음성에서의 음소 인식에 손쉽게 적용될 수 있으나, 신경 회로망은 HMM에 비해 음소 인식의 성능은 저하되지 않지만, 연속 음성에 적용될 경우, 많은 문제점들을 갖고 있으며[5], 현재 이 문제점들에 대한 연구가 활발히 진행되고 있다.

Back propagation 알고리듬이 개발된 이래, 신경 회로망이 많은 음성 인식 시스템에 응용되고 있지만, 이중 몇몇 신경 회로망 음성 인식 시스템만이 음소

특징의 시간 구조를 다루고 있다.

음소 특징 시간 구조를 다룬 신경 회로망중 대표적인 것이 DNN(Dynamic Neural Network)[6], TDNN (Time-Delay Neural Network)[7], TSNN(Time-State Neural Network)[8] 등이 있다. 이중 DNN은 시간 구조를 다루고 있지만 격리 단어에 적용된 것이고, 시간 이동 불변(time-shift invariance) 특성이 불확실하며, TDNN은 시간 이동 불변 특성은 갖고 있지만, 음소 특징의 시간 구조를 상실케 한다.

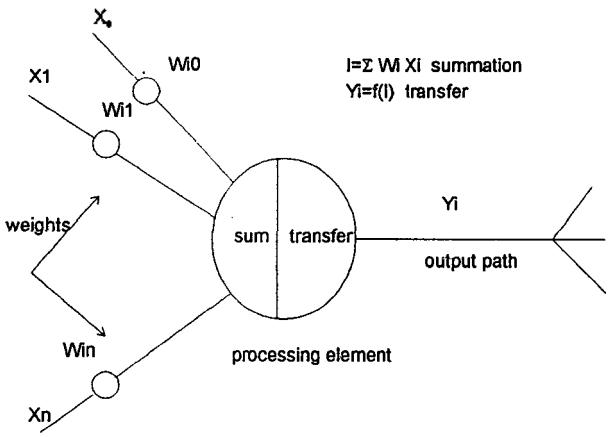
본 논문은 음소 특징의 시간 구조를 다룬 TSNN에 음소 특징의 주파수 구조를 결합시킨 새로운 구조의 신경 회로망을 이용하여 음소 인식을 수행하고자 한다.

Ⅱ. 신경 회로망의 구조와 대표 예

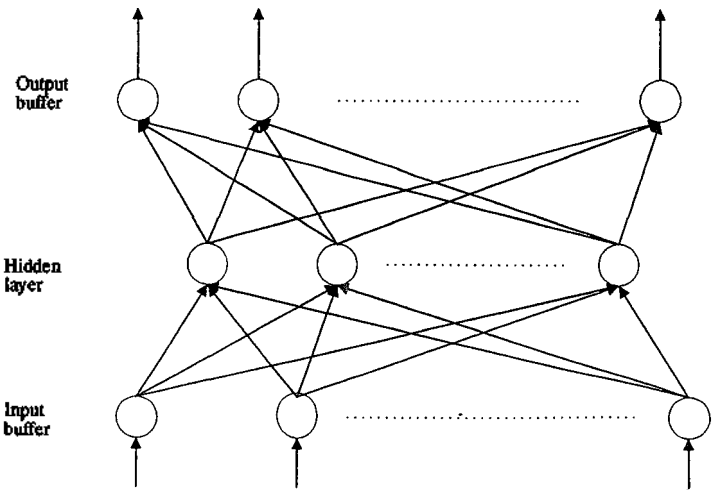
1. 신경 회로망의 기본 구조

신경 회로망에서, 생물학적 신경 세포에 해당하는 것이 처리 인자(processing elements)이다. 이 처리 인자들은 수많은 입력 통로(생물학적 수지상 돌기에 비유)를 갖고 있고, 각 입력 통로에 해당하는 값들을 단순한 덧셈 형식으로 수행하게 된다.

여러 결합된 입력 신호들이 전송 함수(transfer function)에 의해 출력에 연결될 것인가를 결정하게 된다. [그림 1]에 이와같은 신경 회로망의 기본 구성도를 나타내었다.



[그림 1] 신경 회로망의 기본 구성도



[그림 2] 신경 회로망 처리 인자들의 구조

신경 회로망은 위에서 설명한 것과 같이 많은 처리 인자들로 구성되어 있고, 인자들은 [그림 2]에 나타낸 것과 같이 여러 층(multi-layer)으로 분리되어, 각 층의 처리 인자들은 불규칙적으로 서로 연결되어 있다. [그림 2]에서 입력 버퍼(input buffer)는 회로망에 데이터를 입력시키는 작업, 출력 버퍼(output buffer)는 주어진 입력에 대한 회로망의 응답을 유지시키는 작업을 하게 되며, 은닉층(hidden layer)은 입력 버퍼와 출력 버퍼를 서로 구분시켜 주는 역할을 하게 된다.

이와같은 회로망의 작업은 우선 가중치(weight)들을, 각 입력에 대해 원하는 출력을 얻기 위하여, 학습 작업중에 적응시켜야 한다.

2. TDNN(Time-Delay Neural Network)

일반적으로 위와같은 신경 회로망을 이용하여, 음소를 인식하는 방법중 대표적인 것이 TDNN이며, TDNN의 기본 구성도를 [그림 3]에 나타내었다.

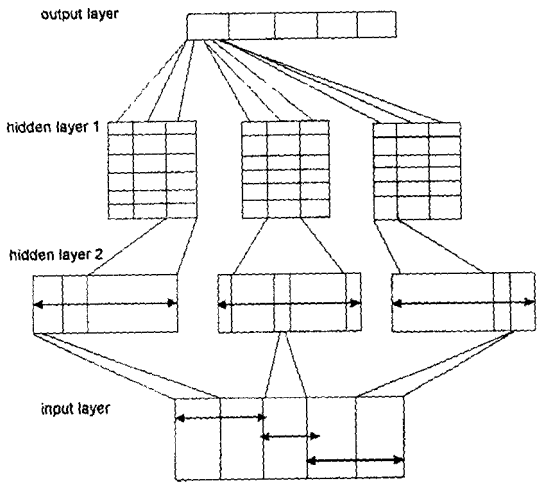
[그림 3]에서 U_i는 입력 인자, Wi는 가중치, Di는 시간 지연을 나타낸다. [그림 3]에 나타낸 것과 같이 TDNN은 은닉 소자로서 지연 소자를 사용하기 때문에, 음성의 시간 변화를 검출해내어 음소들을 판별하는 곳에 이용된다.

이와같은 TDNN은 back-propagation 학습 알고리듬으로 손쉽게 학습시킬 수 있으며, 시간이 이동되고 결속된 가중치 구조에 의해 시간 이동 불변 능력을 갖는 장점을 갖고 있다.

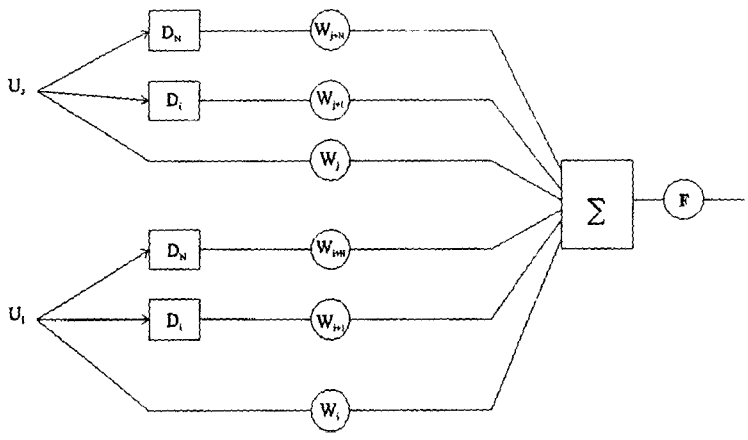
3. TSNN(Time-State Neural Network)

[그림 4]에 TSNN의 기본 구성도를 나타냈으며, 구성도에서 각 층 사이에 있는 window는 서로 결속되어 있고, 시간 이동 불변 특성을 갖는다. [그림 4]의 ⟷ 기호는 window가 이동되는 범위를 나타낸다.

이 TSNN은 3 상태(three state)를 갖게 되며, 각 상태에서 음소 특징을 포착 할 수 있게 된다. 첫번째 상태는 모음으로부터 자음으로 이동되는 음소 특징을, 두번째 상태는 정상 상태에서의 음소 특징을, 그리고 세번째 상태는 뒤에 연결되는 모음 상태로의 천이 상태의 음소 특징을 각각 포착하게 된다.



[그림 4] TSNN의 기본 구성도



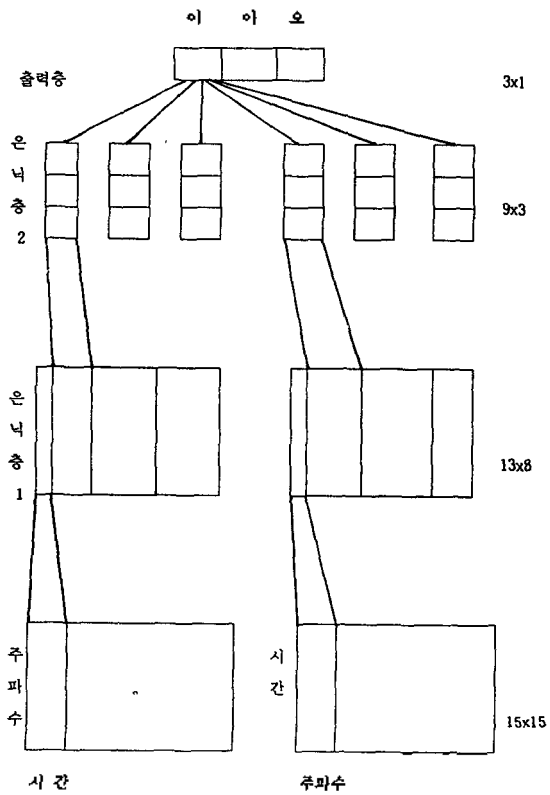
[그림 3] TDNN의 기본 구성도

Ⅲ. 본 논문에서 수행한 신경 회로망

본 논문에서 제안한 음소의 시간과 주파수의 특성을 동시에 다룬 신경 회로망 구조를 [그림 5]에 나타내었다.

일반적으로 TDNN은 back-propagation 알고리듬을 이용하여 손쉽게 학습되어 지고, 시간 이동(time-shifted)되어 함께 연결된 가중치 구조로 부터 시간 이동 불변 특성을 갖는다. 한편 TSNN은 기본적으로 TDNN 구조와 비슷하지만, 출력층에 연결된 가중치들을 세 부분으로 분할시켜, 각 부분이 음소의 각 상태 특성을 대표하게 하여 음소 인식을 수행하게 한다. 즉, 첫번째 부분은 전 음소에서 현 음소로의 시간적인 음소 천이 특성을, 두번째 부분은 현 음소의 정상 특성, 세번째 부분은 현 음소에서 다음 음소로의 시간적인 음소 천이 특성을 나타내어 출력 노드(node)에 연결시켜 주게 된다.

[그림 5]에 나타낸 본 논문에서 수행한 신경 회로망 구조는 음소의 시간적인 변동 특성외에 주파수의 천이 특성을 포함시킨 것이다. 출력 노드에 연결된 은닉층을 살펴보면, TSNN의 시간 천이 특성외에 주파수 변천 특성을 포함시켰다. 즉, 전 음소에서 현 음소로의 시간 천이 특성과 낮은 주파수 대역내에서의 천이 특성이 첫 부분이 되고, 현 음소의 시간 특성과 중간 주파수 대역내에서의 주파수 천이 특성이 두번째 부분이며, 그리고 현 음소로부터 다음 음소로의 시간 특성과 높은 주파수 대역내에서의 주파수 천이 특성이 세번째 부분에서 담당하게 된다.



[그림 5] 본 논문에서 수행한 신경 회로망

이와같이 함으로써 음소의 시간에대한 변화 특성뿐만 아니라, 각 주파수 대역내에서의 변화 특성을 고려하여 음소 인식을 수행할 수 있게된다.

Ⅳ. 실험 결과 및 고찰

1. 실험 데이터

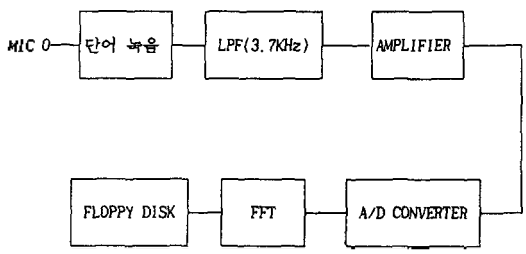
실험 데이터를 구성하는 첫 단계로서 20대 남성 화자를 선택한 후, 숫자음 영(0)부터 구(9)까지를 10번씩 시간 간격을 두어 두번 즉 전체 20번 반복 발음하여, 릴 테이프(reel tape)에 녹음 시켰다.

위와같이 녹음한 후, 이 데이터를 4KHz의 대역폭을 갖는 저대역 통과 필터로 고주파 성분을 제거시켰다.

필터를 통과한 음성 데이터는 16 bit A/D 변환기를 사용하여 디지탈화 시켰으며, 이때 샘플링(sampling) 주파수는 10 KHz로 하였다.

인식 실험에 사용된 데이터는 표준 패턴 구성시 포함되지 않은 음성을 사용하였으며, 인식 파라메터로는 15 채널 주파수 성분을 사용하였다.

이와같은 과정을 [그림 6]에 나타내었다.



[그림 6] 실험 데이터 구성 과정

2. 음소 인식 결과

신경 회로망을 이용하여 음소를 인식하기 위하여 사용한 학습 데이터는 화자 1인이 숫자음을 발생한

것중 각 음소별 5개씩 15프레임을 구분하였으며, 각 프레임을 256 샘플링하여 5msec 주기로 FFT 분석을 수행한 후, 15개의 주파수 선형 대역 에너지를 이용하였으며, 인식할 경우에는 학습시 포함되지 않고 같은 시간에 발생한 모음 15개와 자음 35개, 학습시 사용한 데이터와 다른 시간에 발생한 모음 30개와 자음 70개의 데이터를 사용하였다. 여기에서 FFT 분석 프레임 주기를 5msec로 한 이유는, 무성 자음(ㅅ, ㅊ, ㅍ, ㄱ)의 길이가 모음에 비해 상대적으로 작기 때문에, 무성 자음을 기준으로 하여 프레임 주기를 5msec로 선택하였다.

[표 1]에 일반적인 TDNN 방법을 사용하고, 인식 데이터는 학습시 사용한 데이터와 같은 시기에 발생한 음소를 인식한 결과를 나타냈으며, 인식 결과는 모음 80%, 무성 자음 75%, 유성 자음 100%의 인식률을 얻었다.

[표 2]에 일반적인 TSNN 방법을 사용하고, 인식 데이터는 [표 1]에서 사용한 동일 데이터로 인식한 결과를 나타냈으며, 인식 결과는 모음 100%, 무성 자음 100%, 유성 자음 93.3%의 인식률을 얻어, [표 1]에 나타낸 TDNN 방법보다 TSNN 방법으로 인식한 결과가 향상된 것을 볼 수 있다.

[표 3]에 본 논문에서 제안한 신경 회로망을 사용하고, 인식 데이터는 [표 1]에서 사용한 동일 데이터로 인식한 결과를 나타냈으며, 인식 결과는 모음 93.3%, 무성 자음 90%, 유성 자음 100%의 인식률을 얻었다.

또한, [표 4]에 일반적인 TDNN 방법을 사용하고, 인식 데이터는 학습시 사용한 데이터와 다른 시기에 발생한 음소를 인식한 결과를 나타냈으며, 인식 결과는 모음 70.3%, 무성 자음 80%, 유성 자음 96.6%의 인식률을 얻었다.

[표 5]에 일반적인 TSNN 방법에 [표 4]에서 사용한 동일 데이터로 음소를 인식한 결과를 나타냈으며, 인식 결과는 모음 96.6%, 무성 자음 80%, 유성 자음 83.3%의 인식률을 얻었다.

[표 6]에 [표 3]에서 이용한 신경 회로망에 [표 4]에서 사용한 동일 데이터로 음소를 인식한 결과를 나타냈으며, 인식 결과는 모음 96.6%, 무성 자음 90.2%, 유성 자음 100%의 인식률을 얻었다.

그리고 [표 7]에 인식 방법에 따른 전체 인식률을 나타냈으며, 이 결과 TSNN 방법이 TDNN 방법보다 더 나은 결과를 보였다. 본 논문에서 제안한 방법과 TSNN 방법의 인식 결과, 같은 시기의 데이터에서는 TSNN 방법이 더 나은 결과를 보였지만, 다른 시기의 데이터를 인식한 결과는 TSNN이나 TDNN 방법보다 더 나은 결과를 보여주고 있다. 같은 시기의 인식 결과에서 TSNN 보다 낮은 인식률을 나타낸 것은, 본 논문에서 수행한 같은 시기의 데이터 수가 다른 시기의 데이터 수보다 상대적으로 적기 때문인 것으로 생각된다.

[표 7]에 나타난 것과 같이 일반적인 TDNN방법과 TSNN 방법보다 본 논문에서 제안한 신경 회로망의 인식 결과가 더 나은 것은, TDNN 방법과 TSNN 방법은 시간 성분만 이용하여 인식을 수행하고, 본 논문에서 제안한 신경 회로망은 시간과 주파수 성분을 동시에 입력으로 사용되어 음소의 시간에 따른 주파수 변화뿐만 아니라, 주파수 대역의 상대 변화를 같이 고려하여 인식하기 때문인 것으로 생각된다.

[표 1] TDNN 이용(같은 시기 데이터)

결과 \ 입력	이	아	오
이	5		
아		2	
오		3	5

결과 \ 입력	ㅅ	ㅊ	ㅍ	ㄱ
ㅅ	4		3	1
ㅊ	1	5		
ㅍ			2	
ㄱ				4

결과 \ 입력	ㅇ	ㄹ	ㅁ
ㅇ	5		
ㄹ		5	
ㅁ			5

[표 2] TSNN 이용(같은 시기 데이터)

결과 \ 입력	이	아	오
이	5		
아		5	
오			5

결과 \ 입력	ㅅ	ㅊ	ㅍ	ㄱ
ㅅ	5			
ㅊ		5		
ㅍ			5	
ㄱ				5

결과 \ 입력	ㅇ	ㄹ	ㅁ
ㅇ	4		
ㄹ		5	
ㅁ	1		5

[표 3] 본 논문에서 제안한 신경회로망 이용 (같은 시기 데이터)

결과 \ 입력	이	아	오
이	5		
아		4	
오		1	5

결과 \ 입력	ㅅ	ㅊ	ㅍ	ㄱ
ㅅ	5		1	1
ㅊ		5		
ㅍ			4	
ㄱ				4

결과 \ 입력	ㅇ	ㄹ	ㅁ
ㅇ	5		
ㄹ		5	
ㅁ			5

[표 4] TDNN 이용(다른 시기 데이터)

결과 \ 입력	이	아	오
이	9		
아	1	3	
오		7	10

결과 \ 입력	ㅅ	ㅊ	ㅍ	ㄱ
ㅅ	9	1	3	3
ㅊ	1	9		
ㅍ			7	
ㄱ				7

결과 \ 입력	ㅇ	ㄹ	ㅁ
ㅇ	10		
ㄹ		10	
ㅁ	1		9

[표 5] TSNN 이용(다른 시기 데이터)

결과 \ 입력	이	아	오
이	4		
아		5	
오	1		5

결과 \ 입력	ㅅ	ㅊ	ㅍ	ㄱ
ㅅ	6			2
ㅊ		10		
ㅍ			10	2
ㄱ	4			6

결과 \ 입력	ㅇ	ㄹ	ㅁ
ㅇ	8		
ㄹ		10	3
ㅁ	2		7

[표 6] 본 논문에서 제안한 신경회로망 이용 (다른 시기 데이터)

결과 \ 입력	이	아	오
이	9		
아		10	
오	1		10

결과 \ 입력	ㅅ	ㅊ	ㅍ	ㄱ
ㅅ	10		2	1
ㅊ		10		
ㅍ			8	
ㄱ				9

결과 \ 입력	ㅇ	ㄹ	ㅁ
ㅇ	10		
ㄹ		10	
ㅁ			10

[표 7] 인식 방법에 따른 전체 인식률

인식 방법 / 인식률	TDNN	TSNN	제안한 신경회로망
같은 시기 데이터	84%	98%	94%
다른 시기 데이터	83%	86%	96%

V. 결 론

본 논문에서는 신경 회로망을 이용한 음소 인식에 관한 연구를 수행하였다.

음소를 인식하기 위하여, 본 논문에서는 시간과 주파수 데이터를 신경회로망의 입력으로 하고, 일반적인 TDNN 구조의 출력층에 연결된 가중치들을 세 부분으로 분할하여, 각 부분의 음소의 시간 변화 상태를 각각 담당하게 하여 더 세밀한 음소의 시간 변화

상태를 인식시 이용하였다. 또한 시간 변화 상태외에 주파수 대역의 변화 상태를 음소 인식시 이용하기 위하여, 음소의 주파수 대역을 세 부분으로 나누어 음소 인식을 수행하는 신경 회로망을 구성하였다.

이 결과 본 논문에서 제안한 시간, 주파수 변화를 동시에 이용하여 음소를 인식하는 신경 회로망의 인식 결과는 일반적인 TDNN 방법과 TSNN 방법보다 더 좋은 결과를 얻을 수 있었다. 그리고 학습시 수렴 속도도 일반적인 TDNN 방법보다 본 논문에서 제안한 신경 회로망의 수렴 속도가 더 나은 것으로 나타났다.

앞으로 학습시 인식 대상 음소의 수를 늘려 결과를 살펴보고, 여러 화자의 데이터를 사용하여 화자 독립 음소 인식을 연구한 후, 이 결과를 연속 음성 인식에 적용시킬 예정이다.

참 고 문 헌

1. H.Sakoe, "Two-Level DP matching-dynamic programming based pattern matching algorithm for connected word recognition," IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-27, pp.588-595, Dec. 1979.
2. Y.Linde, A.Buzo and R.M.Gray, "An algorithm for vector quantizer design," IEEE Trans. on Com, Vol. COM-28, Jan., pp.84-95, 1980.
3. R.M.Gray, "Vector quantization," IEEE Mag., Vol. 1, pp.4-29, Jan. 1980.
4. L.R.Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceeding of the IEEE, Vol.77, 1989.
5. R.P.Lipman, "An Introduction to Computationing with Neural Nets," IEEE ASSP Mag., Vol.4, pp. 4-22, 1987.
6. H.Sakoe, et al., "Speaker-independent Word Recognition Using Dynamic Programming Neural Networks," IEEE, ICASSP, pp.29-32, 1989.
7. A.Waibel, et al., "Phoneme Recognition Using Time-Delay Neural Networks," IEEE Trans. Acoust., Speech and Signal Processing, Vol.ASSP-37, March, pp.328-339, 1989.
8. Y.Komori, "Time-State Neural Network(TSNN) for Phoneme Identification by Considering Temporal Structure of Phoneme Features," IEEE, ICASSP, pp.125-128, 1991.
9. Y.H.Pao, Adaptive Pattern Recognition and Neural Networks, Assison-Wesley Pub. Com, 1989.

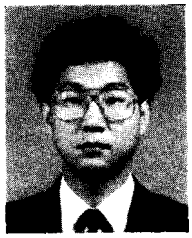
▲李 俊 模(Joon Mo Lee) 1949년 3월 28일생

1975년 2월 : 명지대학교 전자공학과 졸업(공학사)
1979년 8월 : 명지대학교 대학원 전자공학과 졸업(공학석사)
1988년 3월~현재 : 단국대학교 대학원 전자공학과 박사과정 재학중
1980년 3월~현재 : 관동대학교 전자공학과 부교수
※주관심분야 : 멀티프로세서 디자인, 컴퓨터 및 퍼지 제어, 신경 회로망 등임.

▲黃 英 洙(Young Soo Hwang) 1960년 4월 9일생

1982년 2월 : 연세대학교 전자공학과 졸업(공학사)
1984년 2월 : 연세대학교 대학원 전자공학과 졸업(공학석사)
1990년 2월 : 연세대학교 대학원 전자공학과 졸업(공학박사)
1989년 9월~현재 : 관동대학교 전자공학과 조교수
※주관심분야 : 음성인식, 신호처리

▲金 成 鍾(Seong Jong Kim) 1964년 1월 19일생



1987년 2월 : 단국대학교 전자공학과 졸업(공학사)
1989년 2월 : 단국대학교 대학원 전자공학과 졸업 (공학석사)
1993년 3월~현재 : 단국대학교 대학원 박사과정 재학중
※주관심분야 : 병렬처리, Neuro System, Fuzzy, Chaos

▲申 仁 澈(In Chul Shin) 1949년 11월 5일생

1973년 2월 : 고려대학교 전자공학과 졸업(공학사)
1978년 9월 : 고려대학교 대학원 전자공학과 졸업 (공학석사)
1986년 9월 : 고려대학교 대학원 전자공학과 졸업 (공학박사)
1984년~1985년 : 미국 미시간 주립대학교 교환 교수
1979년~현재 : 단국대학교 공과대학 전자공학과 교수
※주관심분야 : 다중처리 컴퓨터, 퍼지제어, 신경 회로망 등임.