Spatio-temporal 방법을 이용한 우리말 모음 인식에 관한 연구

A Study on the Vowel Recognition of Korean Speech using Spatio-temporal Method

송 도 선*, 김 선 일**, 김 석 동***, 이 행 세**
(Do Sun Song*, Sun Il Kim**, Seok Dong Kim***, Haeng Se Lee**)

요 약

본 논문은 신경망을 이용한 우리말 모음에 대한 인식 연구이다. 음성을 나누거나, 음소별 인식이나, 시간 신축 방법을 사용하지 않고 모음을 인식하였다. 시간의 변화에 따른 음성의 변화를 정적인 음성으로 취급하였다. 전처리 과정으로는 단지 모음구간을 검출하여 음성의 앞뒤에 있는 묵음을 제거하였고 모음의 길이를 조사하였다. 10개로 균등히 나눈 프레임에 각 프레임마다 10차의 PARCOR계수를 추출하였다. 신경망의 구조를 간단히 하기 위해서 단모음과 복모음을 구분하여 학습시켰으며, 출력 노드의 수를 감소시키기 위해 이진 코드 형태로 구성하였다.

ABSTRACT

This paper is a study on the Korean vowel recognition using neural network. This system recognizes vowel without performing segmentation, phoneme identification, or dynamic time wrapping. It needs a static pattern approach to recognize a spatio-temporal pattern. The preprocessing only includes preceding and tailing silence removal, and vowel length determination. A 10th-order PARCOR coefficients is extracted for determination on each of 10 equally spaced frames. To simplify a structure of neural network, we trained separately single vowels and diphthongs, and composed binary code form to decrease output nodes.

I. 서 론

음성은 인간이 사용하는 많은 정보 전달 수단 중에서 시각과 함께 가장 자연스럽게 이용되고 있고, 사용하는데 특별한 사전지식이나 별도의 훈련이 필요없는 수단이다. 사회의 정보화가 급속히 진전되면서 인간과 기계와의 접촉이 빈번해 짐에 따라 인간과 기계사이의 의사 전달 방법으로서의 음성의 중요성은 더욱 증대되고 있다.

오늘날의 음성인식기로는 다양한 사람의 연속적인 음성을 인식하기에는 해결해야 할 많은 문제점을 갖고있다. 사람에 따라 음성은 특유의 성질을 갖고 있으며, 시간에 따라 수시로 변화하는 것이 대표적인 문제로 볼 수 있다[1][2]. 어떤 사람은 보통 보다 빠르고/늦게 발음하거나, 크고/작게 말을 하며, 심지어는 같은 사람이라 해도 똑같은 단어를 동일한 형태로 발음하지 않는다[1][3]. 더구나 음성인식에서 올바르게

* 중경공업전문대학 전자계산기과
** 아주대학교 전자공학과
*** 호서대학교 전자계산학과
접수번호 : 1993년 6월 2일

인식이 되었는지 결정하기도 쉽지가 않다. 일반적으로 사람의 경우 듣는 순간 5만개의 단어와 순간적으로 비교해가면서 음성학적이나 구문적이나 의미적인 다양한 분석을 행하면서 인식하고 있으나[1] 아직 음성인식기는 단순한 형태의 분석에 의해 인식되고 있는 실정이다.

지금까지 가장 일반적인 인식알고리즘으로서는 DTW (Dynamic Time Wraping)나 HMM(Hidden Markov Model)이 있다. DTW는 저장되어 있는 표준패턴과 미지의 음성을 동적인 프로그램 기법을 이용하여 정합시키는 방법으로 시간축상에서 표준패턴의 각 부분과 정합시켜 가장 비슷한 것을 찾는 방법이다[5]. 현재까지 가장널리 사용되는 음성인식으로 HMM을 들 수 있다. 음성을 인식하는데 통계적으로 접근하는 방법인 HMM은 음성데이타를 관찰열로 사용하여 HMM에서 만들어진 여러형태의 관찰열 중에서 가장 유사한 것을 선택하는 알고리즘이다[6]. 음소와 음절을 기초로 한 인식이 인공 신경 회로망을 이용하여 여러종류가 발표되어 있다. Kohonen의 자기 구조(self-organizing) 특징도(feature map)는 음소 인식 모델중의 하나이다. 이것은 음소를 인식하기 위한 학습 방법이 자율 학습으로 비슷한 음소끼리 묶어 분류하였다[7]. 음소 인식에 대한 또다른 성공적인 인공 신경 모델은 시간 지연 신경 회로망(Time Delay Neural Network)이다[8][9]. TDNN은 여러층을 갖는 역전파(Back Propagation) 모델을 이용하여 각각의 연속적인 층으로 오랜시간에 대하여 음성을 적용하는 방법으로 가장 위층에서 시간에 불변하는 특성을 가진 음소를 인식하는 방법이다[10].

본 논문의 신경회로망에 사용한 알고리즘으로는 역전파 알고리즘을 이용하였다[11]. 음성인식분야에서 적은수의 어휘를 인식하는데 역전파 모델은 높은 인식률을 얻을수 있다고 발표되어 있다[12]. 그러나 어휘가 많은 음성을 일반적인 역전파 모델을 적용하면 학습시간이 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 재학습시켜야 하는 반복적인 문제가 발생한다. 그러므로 일반적인 역전파 모델로는 많은 어휘에 대해서는 실효성이 없게된다. 이러한 문제점을 보완하기 위해서 본논문에서는 전체인식대상을 몇개의 그룹으로 나누어 학습시키는 방법을 사용하였다[13][14]. 부분망은 그 그룹에 속하는 데이타만을 학습시키므로서 학습시간을 감소시킬 수 있으며 새로운 어휘에 대해서도 손쉽게 확장할 수 있는 장점을 가진다.

Ⅱ. 신경망에 의한 인식

음성은 시간에 따라 수시로 변화한다. 시간에 따라 변화하는 패턴을 신경망으로 인식하기는 쉽지 않다. 일반적인 방법은 음성을 일정한 구간으로 나누어서 시간에 따른 음성의 변화를 최소화하여 사용한다. 본 논문에서는 음성의 시간에 따른 변화를 흡수하기 위해 음성의 전 구간을 10개의 프레임(frame)으로 나누어서 모든 프레임을 한꺼번에 신경망의 입력으로 사용하였다. 본 논문에 사용한 전체적인 방법은 그림 1에 나타내었다.

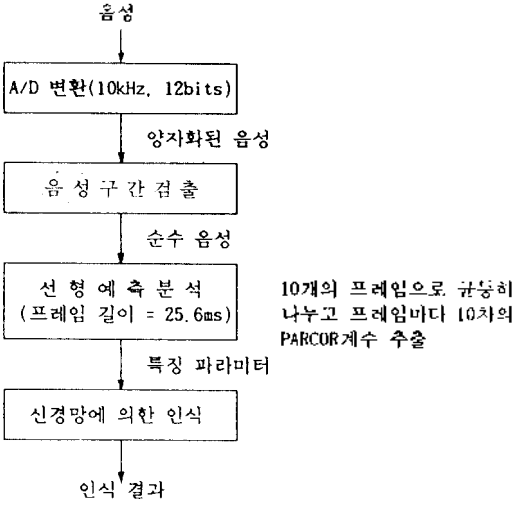

그림 1. 전체 개략도

잡음이 있는 상황에서 음성의 존재 여부를 검출하는 일은 음성처리에 있어서 비교적 어려운 일이다. 컴퓨터에의 인식시 한 단어의 음성이 묵음(Silence)을 앞뒤로 하고 그 사이에 발음 된다는 것을 전제하기 때문에 단어인식을 위해서는 음성 부분과 묵음을 구별해 낼 필요가 있다. 즉 음성의 시작과 끝을 찾는 것을 음성구간 검출이라 한다. 음성 신호의 모음구간 검출은 음성에 해당하는 입력 부분만 처리하도록 함으로써 계산량을 줄이고 인식률을 향상시키는데 중요한 역할을 한다. 정확한 음성구간 검출은 S/N비가 높은 환경에서는 가장 낮은 레벨의 음성(에너지가 낮

은 마찰음)에너지 조차도 배경 잡음에너지를 초과한다. 그래서 단지 에너지만으로도 음성구간을 쉽게 검출할 수 있다. 그러나 대부분의 실제적인 음성 시스템은 15~20 dB 정도의 낮은 S/N비에서 동작된다.

본 논문에서는 프레임내의 평균에너지와 피크 신호 크기를 이용해 음성의 음성구간을 검출 하였다. 음성 신호의 통계적 특성이 시간에 따라 거의 변하지 않는 준 정상적(quasi-stationary)인 사실을 이용해 묵음이 포함된 전음성 구간을 일정한 시간 간격으로 분할(segmentation)하여 프레임(frame)으로 나눈 다음 각 프레임에서 평균 에너지와 최대 신호 크기의 비를 구한다. 평균 에너지와 최대값의 비(rate)가 최대가 되는 프레임을 찾아 양쪽으로 가면서 문턱값(thrshold) 이하가 되는 프레임을 음성의 시작점과 끝점으로 한다. 음성구간을 검출하는 알고리즘을 그림 2에 나타내었다.

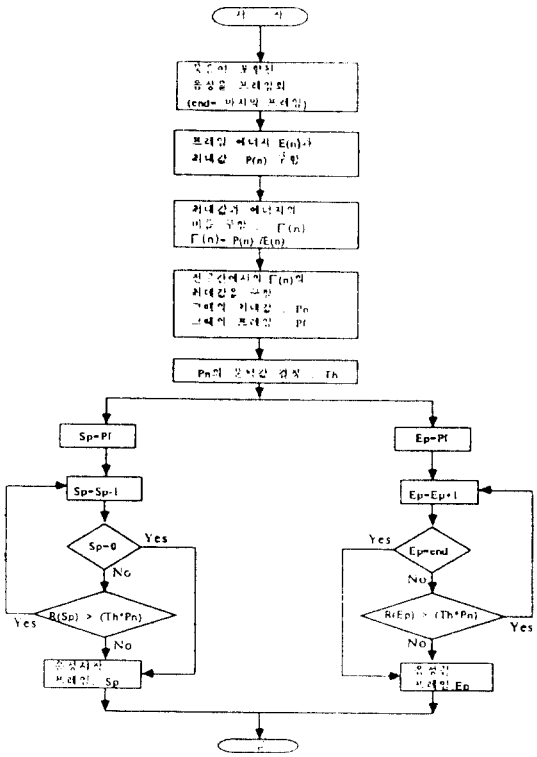

그림 2. 음성구간 검출 알고리듬

각 프레임에서 구한 음성 특징으로는 선형예측분석에 의한 10차의 PARCOR계수를 이용하여 10 frame × 25.6 msec(256 samples)보다 짧게 발음한 음성은 분석 구간을 중첩시키고 길게 발음한 음성은 프레임 사이에 간격을 두었다. 모음 분석을 위해 사용된 성도를 순간적으로 분석할 수 있는 충분히 짧은 시간이다. 또한 각 프레임마다 해밍윈도우(Hamming window)가 사용되었다. 신경망의 입력 갯수는 총 100개(10 frames × 10차 PARCOR 계수)로써 i번째 입력 노드의 의미는

$$i = ((j-1) \times 10) + k$$

여기서 $j = 1, 2, 3, \ldots, 9, 10$

$k = 1, 2, 3, \ldots, 9, 10$

로 표현된다면 j번째 프레임의 k차 PARCOR 계수가 된다.

인식 대상인 모음은 단모음 8개(아·어·오·우·으·이·에·에)와 복모음 12개(야·여·요·유·애·예·와·왜·외·위·위·의)로 총 20개이며, 이 때 출력노드의 수가 20개가 되어야 하나, 본 논문에서는 학습시간및 인식률의 향상을 위해 그림에 나타난 것과 같이 첫째로 단모음과 복모음을 따로 학습시켜서 두 종류의 모음이 혼동되는 것을 피하였다. 단모음과 복모음을 구별하는 것이 선결문제로 대두되는데, 이를 위해서 단모음과 복모음을 구분하는 신경망을 하나 더 추가하였다.

둘째로, 단모음의 인식 신경망의 출력노드를 3개로 줄이고 복모음은 4개로 줄였다. 즉, 출력노드가 N개일 때 binary값은 2^N개가 되므로 이를 이용하여 3개의 출력 노드는 $2^3 = 8$개의 모음을 인식할 수 있고, 복모음은 12개 음성이므로 N을 4로 하였다. 그리고 그림 3에 나타난것 같이 별도로 단모음과 복모음을 구분하는 신경망을 하나 추가하였다. 그림에서 C-net은 음성을 단모음인지 복모음인지 판단하는 신경망으로, 단모음으로 판단된 경우 c1을 활성화시켜 단모음을 인식하는 A-net만을 선택하고, 복모음인 경우는 마찬가지로 B-net을 선택하게 된다. 단모음 인식 신경망의 출력 노드는 이진화로 코드되며 $A_1A_2A_3$가 000, 001, 010, 011, 100, 101, 110, 111은 각각 /아/, /어/, /오/, /우/, /으/, /이/, /애/, /에/를 의미한다. 같은 방법으로 B-net의 $B_1B_2B_3B_4$가 0000 .. 1011은 /야/, .. /의/를 나타낸다.

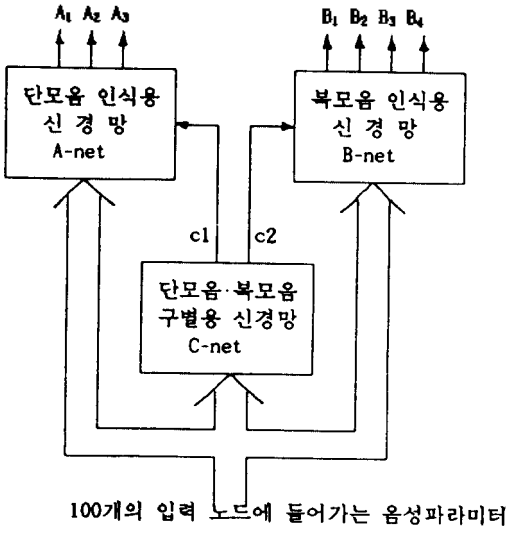

그림 3. 음성인식 신경망의 구조

Ⅲ. 실험및 결과

음성 데이타 수집은 DAT(Digital Audio Tape Recorder)로 방음실에서 다이내믹 마이크(dynamic microphone)를 통해서 약 1.6sec 동안 발음한 음성을 자기테이프(magnetic tape)에 저장하였다. 학습에 사용한 음성은 단모음 8개와 복모음 12개인 총 20종류의 모음을 남성 2사람이 각각 10번씩 발음한 총 400개의 음성(20개 모음×2사람×10번발음)이며, 인식에 사용한 음성은 학습에 참여하지 않은 3사람의 음성으로 10번, 10번, 5번씩 발음한 총 500개의 음성을 대상으로 하였다. 자기 테이프 재생기와 증폭기를 거쳐 나온 음성을 data acquisition board인 DT2801로 음성 크기는 최대치가 ±10V 이내, 샘플링 주파수 10kHz, 양자화 준위 12 비트로 아날로그-디지탈변환(analog to digital conversion)하여 저장하였다. 학습 횟수는 2000회 이내에서 각 음성에 대해 실제의 출력값과 원하는 출력값의 차, 즉 허용오차가 0.1 이내에 들도록 하였다. 실험에서 사용한 음성 데이타 내용과 특징 파라미터를 표 1에 나타내었다.

본 논문에 제안한 방법에 의한 실험결과를 표 2에 나타내었다. 학습기간을 결정하는 계산량을 줄이기 위해 부분학습을 시킨 결과 학습시간이 일반적인 신경망보다 1/4 정도 줄었다. 반복횟수는 시간을 제한하기 위해 2000번이내에서 각음성에 대해 실제의 출력값과 원하는 출력값의 차이, 즉 허용오차가 0.1 이내에 들도록 하였으나, 전체 학습인 경우 1205번에

표 1. 음성 파라미터

대 상 음 성	단모음	/아/, /어/, /오/, /우/, /으/, /이/, /애/, /에/,
	복모음	/야/, /여/, /요/, /유/, /얘/, /예/, /와/, /왜/, /외/, /워/, /위/, /의/
음 성 채 집	학 습	2명의 남성이 발음한 400개 음성
	인 식	3명의 남성이 발음한 총 500개 음성
샘플링 특성		샘플링 주파수 : 10 kHz 양자화 비트수 : 12 bits
분석 구간		프레임 길이 : 25.6 msec 프레임 갯수 : 음성을 균등히 나눈 10개 해밍윈도우 사용
특징 파라미터		100개(10차 PARCOR 계수 * 10개 프레임)

끝났고, 부분학습인 경우 단모음과 복모음을 구분하는 경우가 가장많은 543번에 끝났다.

신경망내의 각 층을 연결하는 가중치(weight)의 수는(입력노드수×중간노드수×출력노드수)이므로 부분학습인 경우 총 18,000개이고, 전체학습의 경우에는 40,000개이다. 계산량을 학습데이타수×가중치수×반복횟수로 나타낼 수 있으므로 부분학습인 경우 총 2.16×10^9번이고 전체학습인 경우 19.3×10^9으로 부분학습이 약1/9정도 감소하였다. 부분학습에 의한 방법에 의해 출력노드수를 감소시킴으로써 가중치의 수가 줄어들었으며 비슷한 데이타로 학습시킴으로써 반복횟수가 감소하며 계산량이 줄어들었다.

표 2. 부분학습과 전체학습에 대한 비교

	부 분 학 습			전체 학습
	단·복구별	단모음인식	복모음인식	
입력노드수	100	100	100	100
중간노드수	20	20	20	20
출력노드수	2	3	4	20
학습데이타수	20*20	8*20	12*20	20*20
가중치 수	4000	6000	8000	40000
반복횟수	543	469	439	1205
학습시간	1h 21m	0h 35m	1h 05m	8h 47m
계산량	0.87×10^9	0.45×10^9	0.84×10^9	19.3×10^9
인식률	99.6%	97.5%	95.7%	94.8%

인식 결과는 표3에 있다. 부분학습일 때의 총 인식률은 95.6%이고, 전체학습의 인식률은 94.8%로 큰 차이는 없었다. 부분학습때의 인식률 계산은 단모음과 복모음 구분이 100%라 할 때의 인식률은 95.8%이나 단모음과 복모음의 구분이 되지 않은 음성이 2개였는데, 하나는 인식률에 영향을 미치는 요소로, /어/가 단모음 인식에서는 인식이 잘되었으나 복모음

으로 판단된 경우이고, 다른 하나는 복모음 인식 신경망에서 /위/가 복모음 /의/로 오인식된 음성이 단모음으로 판단된 경우로 이경우는 총 인식률에는 영향을 끼치지 않았다. 전체 학습 결과에 있는 X는 오인식 결과 다른 음성으로 나타난것이 아니고 출력노드 모두 0가 나타나거나, 1이 두개이상 나타난 경우이다.

표 3. 음성 인식 결과

	인 식 결 과			
	부 분 학 습 (%)		전 체 학 습 (%)	
	인식률(%)	오인식된 음성	인식률(%)	오인식된 음성
아	100		92	와, X
어	96	아	100	
오	92	우(2)	96	X
으	100	오	96	우
이	100		100	
애	100		96	왜
에	100		92	X(2)
야	96	여	96	X
여	96	요	92	워, X
요	96	여	96	여
유	96	요	100	
얘	100		92	왜, 외
예	96	애	84	X(4)
와	92	왜, 워	96	아
왜	96	와	100	
외	100		96	에
워	92	여(2)	88	어, 여
위	84	의(3)	84	의(3), X
의	100		100	
인식률	95.6%		94.8%	

Ⅳ. 결 론

본 연구에서는 우리말의 전체 모음을 대상으로 인식을 하였다. 시간에 따른 음성의 변화를 입력노드수를 확장하여 포함시켰다. 신경망의 구조를 줄이기 위해 출력노드를 이진 코드형식으로 하여 학습 계산량을 감소시켰으며 단모음과 복모음을 구분하여 각각 학습시켜 훈련시간을 줄였다. 전체를 학습할 때와 부분학습때의 인식률은 각각 94.8%, 95.6%로 큰 차이는 없었으나 학습시간은 약 1/4 정도로 부분학습이 감소되었다.

앞으로 자음이 포함된 단어로 확대하여 보다 일반적인 음성을 인식하는 시스템에 대한 연구가 필요하다.

참 고 문 헌

1. T.W. Parsons, *Voice And Speech Processing*, McGraw-Hill Book Company, 1986.
2. R.D. Peacoke and D.H.Graf, "An Introduction to Speech and Speaker Recognition," *Computer*, vol.23, no.8, August 1990.
3. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Engelwood Cliffs, N.J., Prentice-Hall, 1978.
4. R.P. Lippmann, "Review of Neural Networks for Speech Recognition," *Readings in Speech Recognition*, Morgan Kufmann, Inc., 1990.
5. H.Sakoe and S.Chiba, "Dynamic Programing Optimazition for Spoken Word Recognition," *IEEE Trans Acoust., Speech, Signal Processing*. Vol.ASSP-26, pp.43-49, Feb. 1978.
6. Lee, Kai-Fu and H.W. Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM," *Proceedings IEEE ICASSP*. pp.123-126, 1988.
7. T. Kohonen, "The 'neural' Phonetic Typewriter," *IEEE Computer Special Issue on Neural Networks and Neural Computing*, pp.11-22, March 1988.
8. 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.
9. A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, and K.Lang : "Phoneme Recognition : Neural Networks vs. Hidden Markov Models," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, NY, April 1988.
10. A.Noetzel, "Training Hierachical Neural Networks with Temporal Discriminant Cells," *Report to U.S. Army CECOM, Ft. Monmouth*, NJ, October 1990.
11. R.P. Lippman, "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine*, Vol.4, No.2, pp.4-22, April 1987.
12. 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.
13. 김석동, 이행세, "신경망을 이용한 우리말 인식에 관한 연구," 한국음향학회지 제11권 3호, 1992.6.
14. 김석동, 이행세, "복합신경망을 이용한 우리말 단음 인식에 관한 연구," 한국음향학회지 제11권 6호, 1992. 12.

▲송 도 선(正會員) 1956年 4月 14日生

1977年 : 영남대 전자공학과 졸업
1981年 : 고려대학 대학원 전자공학과 졸업
1990年 : 아주대학교 박사과정 수료
1981年 ~ 현재 : 중경공업전문대학 전자계산기과 부교수
※주관심분야 : 인공지능 및 신호처리

▲金 錫 東 1957년 8월 23일생

1982년 2월 : 아주대학교 전자공학과(공학사)
1984년 2월 : 아주대학교 대학원 전자공학과(공학석사)
1993년 2월 : 아주대학교 대학원 전자공학과(공학박사)
1984년 3월 ~ 현재 : 호서대학교 전자계산학과 부교수
※주관심분야 : 음성인식, 신경회로망, 음성신호처리

▲金 善 一(正會員) 1960年 3月 19日生

1983年 2月 : 아주대학교 전자공학과 졸업(공학사)
1985年 2月 : 아주대학교 대학원 전자공학과 졸업(공학석사)
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 ~ 현재 : 거제 전문대학교 전자과 조교수
1990年 8月 ~ 현재 : 아주대학교 박사과정 재학중
※주관심분야 : 음성 및 신호처리, 패턴인식, Robot Vision 등임

▲李 幸 世(正會員) 1943年 8月 29日生

1966年 : 전북대학교 전기공학과 공학사
1972年 : 서울대학교 전자공학과 공학석사
1984年 : 고려대학교 전자공학과 공학박사
1968年 ~ 1970年 : 해군사관학교 전자공학 교관
1973年 ~ 현재 : 아주대학교 교수
1982年 ~ 1983年 : 미국 Columbia Univ.(N.Y.) 객원교수
1987年 ~ 1988年 : 프랑스 INRIA(Paris) 객원교수
1992年 ~ 현재 : 기제전문대학 학장
1973年 ~ 현재 : IEEE 회원
※주관심분야 : 문자 및 음성인식, 인간 기계 인터페이스, 인공지능, 신경회로망, VLSI 회로, 디지탈 신호처리 등임