음성 인식 신경망을 위한 음성 파라미터들의 성능 비교

A Comparative Study of Speech Parameters for Speech Recognition Neural Network

김 기 석* 임 은 진* 황 희 융*

(Kiseok Kim* Eunjin Im* Heeyeung Hwang*)

요 약

음성 인식에 신경망 모델을 적용하는 많은 연구들이 있었지만, 주된 관심은 음성인식에 적합한 구조와 학습 방법이었다. 그러나 음성인식에 신경망 모델을 적용한 시스템의 효율 향상은 모델 자체의 구조뿐 아니라, 신경망 모델의 입력으로 어떤 음성 파라미터를 사용하는가에 따라서도 큰 영향을 받는다. 본 논문은 기존 음성인식에 신경망 모델을 적용한 많은 연구들에서 사용한 음성 파라미터를 살펴보고, 대표적인 음성 파라미터 6개를 선정하여, 같은 데이타와 같은 신경망 모델 하에서 어떻게 성능이 달라지는지를 분석한다. 인식 실험에 있어서는 한국어 파열음 9개에 대한 8개 데이타 집합과 모음 8개에 대한 18개 데이타 집합을 음성 파라미터로 하고 신경망 모델은 순환 신경망 모델을 사용하여 노드의 수를 일정하게 한뒤 다양한 입력 파라미터의 성능을 비교하였다. 그 실과 선형 예측 계수로부터 얻어진 delta cepstrum의 음성 파라미터가 가장 좋은 성능을 보였으며 이때 인식률은 같은 학습 데이타에 대해 파열음 100.0%, 모음 95.1%이었다.

ABSTRACT

There have been many researches that uses neural network models for automatic speech recognition, but the main trend was finding the neural network models and learning rules appropriate to automatic speech recognition. However, the choice of the input speech parameter for the neural network as well as neural network model itself is a very important factor for the improvement of performance of the automatic speech recognition system using neural network. In this paper we select 6 speech parameters from surveys of the speech recognition papers which uses neural networks, and analyze the performance for the same data and the same neural network model. We use 8 sets of 9 Korean plosives and 18 sets of 8 Korean vowels. We use recurrent neural network and compare the performance of the 6 speech parameters while the number of nodes is constant. The delta cepstrum of linear predictive coefficients showed best result and the recognition rates are 95.1% for the vowels and 100.0% for plosives.

I. 서 론

컴퓨터가 사용되기 시작한 이래로 컴퓨터는 대량의 계산을 고속 처리하는데 있어서 인간보다 월등히 능력을 보여왔지만 음성의 인식이나 화상의 인식 등 인간의 인지를 흉내내는 영역에 있어서는 아직까지 그 능력이 인간에 훨씬 못 미치고 있다. 이는 인간의 인식 메카니즘과 현재의 컴퓨터의 기본 구조가 되고 있는 내장형 프로그램 방식의 처리 방식의 차이점에서 기인한 것으로 보여진다. 따라서 이러한 응용 영역에 대하여 신경 생리학 분야에서 McCulloth and Pitts[1]가 연구한 뇌와 신경의 모델로부터 발전되어진 신경망 모델을 적용한 연구가 활발히 진행되고 있다. 그러나 이러한 접근 방법도 확실한 성능 향상을 보이고 있지 않다.

* 서울대학교 공과대학 컴퓨터공학과
접수일자 : 1992. 4. 27.

신경망을 통한 성능 향상을 위한 연구 방향은 다음과 같다. 첫째로 신경망 모델의 구조나 학습 방법의 연구이다. 이에 관한 연구는 정적, 동적 음성 패턴에 적합한 많은 모델들이 제안되고 있으며 또한 기존 모델과의 결합방법들도 제시되고 있다[2]. 둘째로 문장이나 대화 단위로부터 얻어지는 문법적, 의미론적 정보를 이용하여 음소, 단어 단위의 인식률을 높이는 방법에 대한 것이 있다[3].

그러나, 신경망 모델 자체와 무관하게, 좀더 정확한 인식을 위해 신경망 모델에 어떤 특징 파라미터를 사용하는 것이 유리한가에 대한 입력되어야 하는지를 연구도 중요하다. 즉 신경망을 이용한 음성 인식 분야에서 더 좋은 성능을 얻기 위하여 연구해야 할 분야로는 신경망의 입력으로 음성의 특징을 가장 잘 표현할 수 있는 파라미터가 무엇인자를 찾아내는 것이다.

본 연구의 목적은 앞으로의 신경망을 이용한 한국어 음성 인식 연구를 위해 현재까지 알려진 특징 파라미터 중 가장 적합한 음성 파라미터를 선정하기 위하여 기존의 신경망을 이용한 음성 인식 연구에서 사용된 여러가지 음성 파라미터를 조사한 후, 중요한 6가지의 음성 파라미터를 같은 데이타 셀과 동일한 신경망 모델(순환 신경망 모델)을 이용한 적용하여 파열음과 모음의 인식 및 학습에 적용하여 인식률 및 학습 속도 등을 비교함으로써, 상대적으로 좋은 파라미터가 무엇인지를 조사한다.

Ⅱ. 음성 파라미터

2-1. 음성 파라미터의 선정

비교할 대상이 될 음성 파라미터를 선정하기 위하여 기존의 신경망을 이용한 음성인식 실험에 사용된 음성 파라미터들을 조사하였다. 표 1에는 기존에 신경망을 이용한 음성 인식 연구에서 사용된 음성 파라미터들의 사용예와 빈도수를 나타낸다. 기존 음향학적 실험에서의 음성 파라미터 사용 경향은 크게 pitch period, 포만트, 에너지 등의 음향학적 파라미터를 추출하여 사용하는 경향과 filterbank 계수, 선형 예측 계수, 이로부터 계산된 cepstrum 계수와 같은 파라미터들을 사용하는 경향으로 나누어 볼 수 있고 상당수의 인식 실험들에서는 이러한 음성 파라미터들을 2가지 이상을 동시에 사용하고 있었다.

이를 참고로 하여 본 실험에서는 mel scale filterbank 계수와 인접한 filter bank 대역들이 반씩 겹치도록 한 overlapped mel scale filter bank의 계수, 이 계수들의 대수 에너지로부터 얻은 cepstrum의 파라미터들과 선형 예측 계수들을 대상으로 선정하였다. 또한 앞으로의 음성 파라미터 사용 경향으로 볼 때 가능한 한 인간의 청각 신경계가 받아들이는 파라미터와 유사한 파라미터를 사용하려는 경향을 고려하여 auditorily based spectral tranform으로 제안된 변형된 Fourier-t-transform으로부터 얻어지는 계수를 추가한 이 6가지 파라미터에 대하여 인식 실험을 실시하고 그 결과를 비교하였다.

표 1. 음성인식 신경망 모델 연구에서 주로 사용되는 음성 파라미터들

사용된 파라미터	사용된 실험 수
raw signal	1
acoustic features	14
autocorrelation	1
(mel scaled) filterbank	9
linear predictive coeff.	7
perceptual linear prediction coeff.	2
LPC cepstrum	8
mel scale filterbank cepstrum	11
delta mel cepstrum	2
auditory model	4
위의 2가지 이상의 결합	15

2-2. 본 연구에서 사용한 음성 파라미터

본 논문에 비교되는 음성 파라미터들은 cutoff frequency 4.4 kHz로 lowpass filtering된 음성 신호를 10.4 kHz로 샘플링하여 얻은 음성 데이타에 대하여 256개의 샘플을 한 프레임으로 하여 매 6.4 ms마다 다음과 같은 처리를 행하여 신경망의 입력으로 사용한다.

파라미터의 값의 차이를 중간부분에서 강조하기 위하여 추출된 파라미터 값은 전체적인 크기 순으로 볼 때 작은 20%의 값은 0.0, 큰 20%의 값은 1.0, 그 중간 값에대해서는 선형적으로 정규화한다.

(1)mel scale filter bank 계수

인간의 청각 메카니즘이 낮은 주파수 대역에 대하여는 높은 해상도를 갖고 높은 주파수 대역으로 갈수록 해상도가 더 낮아짐을 고려하여 낮은 주파수 대역

에서는 선형적이고 높은 주파수 대역에서는 대수적인 간격을 갖는 scale로써 mel scale[4]이나 bark scale[5]이 제안되었다.

이에 따라 각 frame을 Fourier Transform 한 결과를 29개의 Mel scale의 filter bank로 나누어 각 대역의 에너지를 계산하였다.

이 때, Mel scale과 주파수와의 관계는 $M=2595\log_{10}(1+f[Hz]/700)[mel]$[6]의 근사치로 계산되었으며 각 대역은 83 mel 간격을 가진다.

(2)Overlapped mel scale filter bank 계수

각 frame을 Fourier Transform 한 결과를 각 대역이 서로 겹치는 부분이 있도록 56개의 Mel scale의 filter bank로 나누어 각 대역의 에너지를 계산하였다. 이 때 각 대역의 간격은 83 mel이고 겹치는 부분은 41.5 mel이다.

(3)Overlapped mel scale filter bank 계수의 Cepstrum(MFCC)

관찰되는 신호를 주파수 영역으로 변환하고 여기서 위상 성분을 제거한 후 역변환한 계수를 cepstrum 계수라 한다[7].

본 실험에서는 다음의 관계식[8]에 의하여 (2)에서 구해진 overlapped mel scale filter bank 계수의 대수 에너지를 역변환하여 cepstrum 계수를 추출하여 신경망의 입력으로 사용한다.

$$MFCC_i=\sum_{k=1}^{20} X_k \cos(i(k-1/2)\pi/20),\ i=1,\ \ldots,\ 16$$

여기서 X_k는 k번째 filter bank의 대수 에너지 출력이다.

(4)선형 예측 계수(LPC)

이전 p개의 음성 샘플의 선형 결합에 의하여 다음 음성 샘플을 예측할 수 있다는 가정 하에 예측된 음성 샘플과 실제의 음성 샘플 사이의 오차를 최소화하는 계수들을 음성 파라미터로 사용할 수 있다[7].

이러한 조건을 만족하는 선형 예측 계수 α_k를 찾기 위한 방법으로는 autocorrelation method나 covariance method[9]들이 있으며 이 방법들의 해를 구하는 알고리즘으로 Gram-Schmidt 알고리즘이나 Levinson-Durbin 알고리즘[10] 등이 제안되었다.

본 실험에서는 Levinson-Durbin 알고리즘에 의하여 12차의 선형 예측 계수를 구하여 음성 파라미터로 사용한다.

(5)선형 예측 계수의 delta cepstrum

(4)에서 계산된 12차 선형 예측 계수로부터 다음 관계식에 의하여 LPC cepstrum을 구하고[9] 이와 함께 현재 frame의 LPC cepstrum과 3 frame 이전의 LPC cepstrum과의 차이를 계산하여 함께 신경망의 입력으로 사용한다.

$$LPCC_i=LPC_i+\sum_{k=i}^{i-1}(k-i)/i\ LPCC_{i-k}\ LPC_k\ (LPC_i:i\text{번째 LPC 계수})$$

(6)변형된 Fourier-t-transform계수 (FTT 계수)

M. Beham은 문헌[10]에서 인간의 청각 해상도가 주파수 해상도에 있어서 고주파수 대역에서보다 저주파수 대역에서 더 높고, 시간 해상도에 있어서는 저주파수 대역에서보다 고주파수 대역에서 더 높은 점을 반영하는 spectral transform으로서 변형된 Fourier-t-transform을 제안하고 이의 순환적 구현 알고리즘을 제시하였다. 본 실험에서는 이 알고리즘에 의해 매 6.4 ms마다 계산된 29개의 계수를 신경망의 입력으로 사용하였다.

Ⅲ. 신경망 모델의 구조와 학습 데이타

실험에 사용된 신경망의 구조는 그림 1과 같이 첫번째 은닉층과 출력층에 대하여 문맥층을 두는 변형된 순환 신경망 구조를 가지며 입력 윈도우 크기는 4이므로 248(4×62)개의 입력 노드, 15개의 첫번째 은닉 노드, 30개의 두번째 은닉 노드, 모음의 경우 8개, 파열음의 경우 9개의 출력 노드, 30(2×15)개의 첫번째 문맥 노드, 모음의 경우 16(2×8)개, 파열음의 경우 18(2×9)개의 두번째 문맥 노드를 가진다.

신경망의 구조를 간계하기 위하여 입력 피라미터의 차원이 작은 경우는 이전 파라미터의 값을 복사하여 입력한다.

신경망의 학습은 오류 역전파 알고리즘을 사용하며 학습률은 0.1, 문맥층으로부터 은닉층으로의 연결 가중치 학습률은 0.1, bias의 학습률은 0.1을 사용하였다.

학습 데이타 셋은 1인의 남성 화자가 발음한 "ㅏ/ㅓ/ㅗ/ㅜ/ㅡ/ㅣ/ㅐ/ㅔ" 8개의 모음 세트 16개와 "ㄱ/ㄷ/ㅂ/ㄲ/ㄸ/ㅃ/ㅋ/ㅌ/ㅍ" 9개의 파열음 세트 8개이며 파열음에 대한 신경망과 모음에 대한 신경망을 따로 각각 500회까지 학습시켰다.

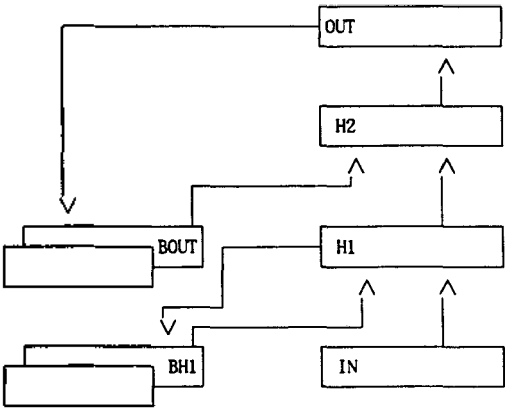

그림 1. 본 연구에서 사용한 순환 신경망 모델

Ⅳ. 실험 결과와 비교 분석

6개의 음성 파라미터를 500회까지 학습시킬 때 학습 셀에 대한 인식률의 수렴 속도를 모음과 파열음에 대하여 각각 도시하면 그림 2,3, 그림 4,5와 같다. 인식 실험데이타는 학습 데이타를 사용하였다.

500회 학습후의 인식률만을 표 2에 표시하였다. 이 결과 선형 예측 계수로부터 얻어진 delta cepstrum이 모음 95.1%, 파열음 100.0%의 가장 높은 인식률을 보였다. 그림에서 TSS는 total sum of squared error, r. rate는 인식률을 나타낸다.

실험 결과 모음에 있어서는 선형 예측 계수를 제외한 나머지 5개 파라미터에서는 거의 비슷한 수렴 속도와 인식률을 보이고 있음을 알 수 있다. 이는 모음은 주로 그 조음 위치에 따라 구별할 수 있는데 각 조음 위치에 따라 제 1, 제 2포만트의 값이 특징적으로 분포하기 때문에 주파수 영역에서 얻어지는 파라미터들에서는 쉽게 구별이 되기 때문인 것으로 보여진다.

파열음의 인식 실험에서 mel frequency cepstrum 계수나 LPC 계수의 delta cepstrum이 월등한 인식율을 보이고 있는데 이는 파열음의 조음에 있어서 excitation 함수는 noise 함수로 각 파열음 간에 차이를 보이지 않는 반면 그들 간의 차이를 주는 성분은 성도 전달 함수로서 cepstrum 계수가 성도 전달 함수와 excitation 함수를 분리하여 제시하여 주기 때문인 것으로 추측된다. Delta cepstrum이 좋은 결과를 내는 것으로부터 음성 파라미터의 시간적인 변화 역시 중요한 단서가 됨을 알 수 있다. 또한 mel scale

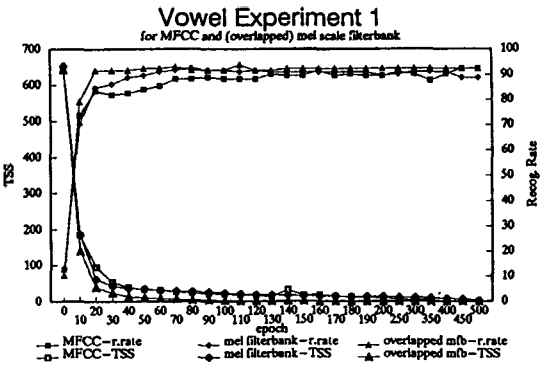

그림 2. mel frequency cepstrum 계수 (MFCC), mel scale filterbank, overlapped mel scale filter bank 음성 파라미터의 모음에 대한 인식률의 수렴 속도

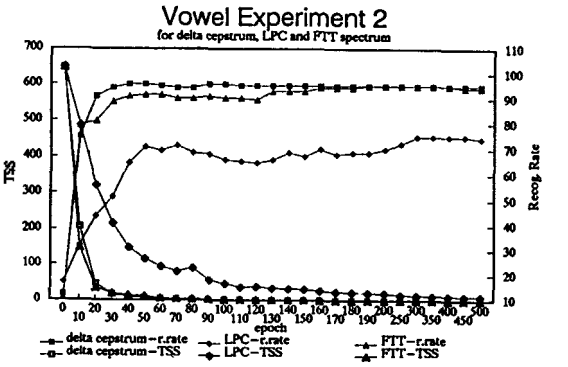

그림 3. delta cepstrum 계수, 선형 예측 계수 (LPC), 변형된 Fourier-t-transfom 계수 (FTT) 음성 파라미터의 모음에 대한 인식률의 수렴 속도

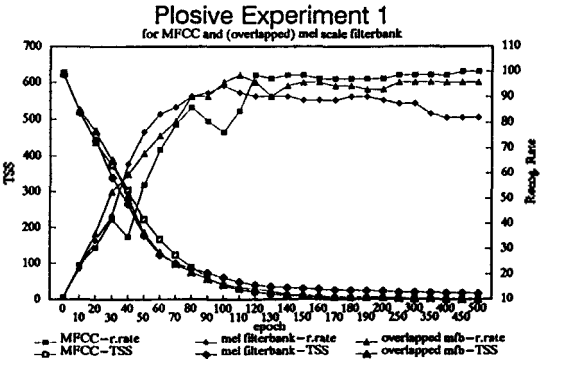

그림 4. mel frequency cepstrum 계수 (MFCC), mel scale filterbank, overlapped mel scale filter bank 음성 파라미터의 파열음에 대한 인식률의 수렴 속도

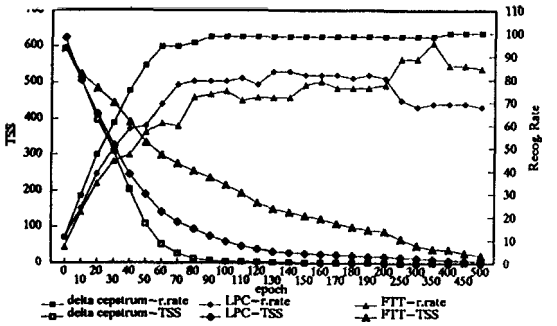

그림 5. delta cepstrum 계수, 선형 예측 계수 (LPC), 변형된 Fourier-t-transfom 계수 (FTT) 음성 파라미터의 파열음에 대한 인식률의 수렴 속도

표 2. 500회 학습 후의 각 파라미터의 인식률

인식률 / 파라미터	모 음	파 열 음
Mel scale filter bank	88.9%	81.9%
Overlapped mel scale filter bank	92.4%	95.8%
Cepstrum of overlapped mel scale filter bank	92.4%	100.0%
선형 예측 계수	74.3%	68.1%
Delta cepstrum of LPC	95.1%	100.0%
Auditorily based spectral transformation(FTT)	94.4%	84.7%

filterbank와 overlapped mel scale filterbank의 학습 곡선에서 청각 신경의 분포와 비슷한 overlapped mel scale filterbank가 mel scale filter bank보다 음성의 특성을 더 잘 반영하고 있다고 추측할 수 있다. 파열음에서도 역시 선형 예측 계수는 인식률이 좋지 못한데 이는 선형 예측 계수가 전이적인 특성보다는 다소 안정적인 (steady) 특성을 가지는 음소에 더 적합한 계수임을 보여주는 결과로 해석할 수 있다. 마지막으로 FTT 계수의 인식률 곡선을 보면 이미 350회 학습 후에 약 95%까지 올라 갔다가 인식률이 낮아진 것으로 보아 아직 학습이 수렴하지 못하고 진동하고 있는 중이기 때문인 것으로 짐작되며 그 이후의 학습에서는 더 많은 인식률의 향상이 기대된다.

V. 결론 및 앞으로의 연구 방향

한국어 모음과 파열음 셀에 대하여 filter bank output, overlapped filterbank output, mel frequency cepstrum 계수, LPC 계수, delta cepstrum of LPC, 변형된 Fourier-t-transfom 계수의 6가지의 음성 파라미터를 추출하여 순환 인공 신경망 모델에 학습시켜 인식률과 수렴 속도를 비교하였다.

사용한 데이타 세트는 1인의 남성화자가 발음한 "ㅏ/ㅓ/ㅗ/ㅜ/ㅣ/ㅐ/ㅔ" 8개의 모음 세트 18개와 "ㄱ/ㄷ/ㅂ/ㄲ/ㄸ/ㅃ/ㅋ/ㅌ/ㅍ" 9개의 파열음 세트 8개를 분리하여 각기 따로 학습시켰다. 비교 결과 delta cepstrum of LPC가 500번의 학습 후에 모음 95.1%, 파열음 100.0%의 가장 높은 인식률을 보였다.

앞으로의 음성 파라미터에 관한 연구에서는 좀더 인간의 청각 신경계에서 자극을 받아들이는 형태와 유사한 파라미터를 찾아내는 것과 잡음에 강한 음성 파라미터를 찾아내는 것, 화자에 따른 변이에 무관한, 또는 반대로 화자 인식에 이용될 수 있도록 화자의 특성을 잘 나타내는 파라미터를 제안하는 것이 필요할 것이다. 또한 이러한 파라미터를 신경망의 입력으로 사용할 때 성능 개선을 위하여 일정 범위의 값으로 정규화하는 방법의 연구도 이루어질 것이다.

음소 인식 단계에서 더 좋은 인식률을 얻기 위하여는 인간의 청각 신경계를 연구하여 이와 유사한 파라미터를 찾아내는 방법이 있을 것이다. 그러나, 인간의 음성 인식 실험에서도 알 수 있듯이 음소 인식 단계에서 100%의 인식률을 기대할 필요는 없는 것으로 보여진다. 왜냐하면 인간은 개개의 음소 인식률이 완전하지 않더라도 문맥 등의 상위 차원 정보를 이용하여 문장을 이해하고 있기 때문이다. 따라서 음성 파라미터의 연구는 인간과 같이 상위 차원의 정보를 이용하는 연구와 병행되어야 할 것이다.

또한 이 실험은 고립 음소를 인식하는데 국한되었으나 연속된 문장 상에서 음소나 단어의 위치를 발견하고 이를 인식하는 연구가 음성 인식의 실용적인 응용을 위하여 필수적으로 요구된다.

참 고 문 헌

1. McCulloch, W. S. and W. Pitts(1943) A Logical Calculus of Ideas Immanent in Nervous Activity *Bulletin of Mathematical Biophysics* 5, pp.115-133.

2. L. T. Niles and H. F. Silverman, "Combining hidden Markov models and neural network classifiers," Proc. ICASSP, April 1990, pp.417-420.

3. R. B. Allen, "Several studies on natural language

and back propagation," Proc. ICNN, June 1987, vol. II.

4. Beranek, L. L. (1949) Acoustic Measurements, New York. p.914.
5. E. Zwicker and E. Terhardt (1980) "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," J. Acoust. Soc. Am. 68, pp.1523-1525.
6. Douglas O'Shaughnessy (1987) Speech Communications-Human and Machine.
7. Schafner, R. W. and Rabiner, L. R. (1975) "Digital representation of speech signals," Proc. of IEEE 63 (4), pp.662-667.
8. Davis, S. B. and Mermelstain, P. (1980) "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. on ASSP, ASSP-28, pp.357-366.
9. Markel, J. D. and Gray, A. H. Jr. (1976) "Linear prediction of speech," Springer-Verlag.
10. Beham, M (1991) "An auditorily based transformation of speech signals," proc. of eurospeech 91, vol.3, pp.1437-1440.

▲김 기 석

1984년 : 서울대학교 공과대학 전자계산기공학과 (학사)
1987년 : 서울대학교 공과대학 컴퓨터공학과 (석사)
1987년－현재 : 서울대학교 공과대학 컴퓨터공학과 박사과정
연구분야 : Speech Recognition System, Artificial Neural Network, Pattern Recognition and AI, Multimedia System.

▲황 희 웅

1964년 : 서울대학교 공과대학 전기공학과 졸업
1974년 : 동 대학원 전기공학과 박사
현재 : 서울대학교 컴퓨터공학과 교수
연구분야 : Microcomputer, Artificial Neural Network, Speech and Character Recognition.

▲임 은 진

1991년 : 서울대학교 공과대학 컴퓨터공학과 (학사)
1992년－현재 : 서울대학교 공과대학 컴퓨터공학과 석사과정
연구분야 : Artificial Neural Network, Speech Recognition System.