VQ와 Multi-layer perceptron을 이용한 단모음 인식에 관한 연구

A Study on Single Vowels Recognition using VQ and Multi-layer Perceptron

안 태 옥*, 이 상 훈**, 김 순 협*

(Tae Ok An*, Sang Hoon Lee**, Soon Hyob Kim*)

요 약

본 논문은 불특정 화자의 단모음 인식에 관한 연구로써, VQ(Vector Quantization)와 MLP(multi-layer perceptron)에 의한 음성 인식 방법을 제안한다. 이 방법은 VQ codebook을 구하고 이를 이용해서 관측열(observation sequence)을 구해 각 codeword가 데이타로부터 가질 수 있는 확률값을 계산하여 이 값을 신경 회로망의 입력으로 사용하는 방법이다.

인식 대상어로는 한국어 단모음을 선정하였으며 10명의 남성 화자가 8개의 단모음을 10번씩 발음한 것으로 시스템의 효율성을 알아보기 위해 VQ/HMM(hidden Markov model)에 의한 인식과 비교 실험한다. 실험 결과에 의하면, 시스템의 단순성에도 불구하고 학습능력이 뛰어난 관계로 VQ/HMM보다 VQ와 MLP에 의한 음성 인식이 인식률이 향상됨을 보여준다.

ABSTRACT

This paper is a study for speaker-independent recognition of single vowels, and we propose a method of speech recognition using VQ(vector quantization) and MLP(multi-layer perceptron). This method makes a VQ codebook, obtains the observation sequence using VQ codebook, calculates the probability value by comparison between each codeword and the data, and then uses these probability values as the input value of the neural network.

Korean signle vowels are selected for recognition experiment, and this recognition experiment, which is performed by ten times utterances of eight single vowels pronounced by ten male speakers, is compared with the recognition experiment by VQ/HMM(hidden Markov model) to investigate the efficiency of the system. According to the experimental result, it is shown that recognition rate of speech recognition by VQ and MLP is better than that of VQ/HMM because of its excellent learning ability in spite of its simplicity.

I. 소 개

음성은 인간의 가장 자연스러운 통신 방법으로, 인간과 기계 사이의 자연스러운 통신을 위해 음성 인식에 대한 연구가 꾸준히 진행되어 왔다. 사람의 음성을 인식하기 위한 알고리즘은 DTW(dynamic time warping)[1], VQ(vector quantization)[2], HMM(hidden Markov model)[3,4] 등을 이용하여 여러 가지 방법으로 연구되어 왔는데 그 중에서도 HMM은 확률론을 도입한 알고리즘으로 1980년대에 다른 방법보다 우수한 장점이 있는 관계로 많은 연구가 이루어지고 있다. 그러나, 이 HMM도 연속 음성, 무한대어휘, 화자독립이라는 음성 인식의 세가지 최종 목표를 해결하는데 여러 문제점이 지적됨에 따라 새로운 방법에 의한 음성 인식 방법이 대두되었는데 그것이 바로 인공 신경 회로망을 이용한 음성 인식에 관한 연구이다.

이 신경 회로망이란 인간의 두뇌가 대량의 복잡한

*광운대학교 전자계산기 공학과
**광운대학교 부설 전산원
접수일자 : 1992년 12월 28일

데이타를 병렬 처리할 수 있다는 사실에 근거하여 새로운 계산 방식 및 패턴 인식의 방법으로 음성 인식에서는 TDNN(time delay neural net)[5], LVQ(learning vector quantization)[6]등 여러 가지 방법으로 적용되어 연구되고 있다.

그러나, 본 연구에서는 새로운 방법의 VQ와 MLP에 의한 인공 신경 회로망을 이용한 음성 인식 방법을 제안한다. 이 VQ와 MLP에 의한 음성 인식 방법은 이산 HMM에서 VQ codebook을 이용하여 관측열(observation sequence)을 구하고 이를 이용하여 HMM 파라메타(parameter)들을 최적화 시키는 것과는 달리, VQ codebook의 codeword와 각 토큰(token)의 프레임과의 거리값을 계산하여 관측열을 구한 후 각각의 codeword가 가질 확률값을 계산하고 이를 입력 특징으로 하였으며 학습 알고리즘으로는 Rumelhart등이 제안한 back propagation 알고리즘을 이용하였다.

본 연구는 제안된 VQ와 MLP에 의한 신경 회로망을 이용한 음성인식 방법 이외에도 비교를 위해 최근까지도 많이 이용되고 있는 인식 방법인 HMM에 의한 음성 인식과 비교한다. 여기에서 VQ codebook을 작성하는데 사용된 vector는 10차 LPC cepstrum 계수이고, 대상 어휘는 8개의 모음이며, 화자는 10명의 남성으로써 세화자가 5번 발음한 음성으로 학습시켰으며, 그 화자가 발음한 나머지 발음 5번을 포함하여 10명의 화자가 발음한 800개의단모음을 인식 실험하였다.

Ⅱ. 이 론

1. VQ(vector quantization) codebook 작성

VQ란 벡터의 sequence를 통신이나 디지탈 채널에 저장하기에 적당한 디지탈 sequence와 mapping하기 위한 시스템이다. VQ의 가장 큰 목적 중 하나는 데이타 압축으로 데이타의 신뢰성을 잃지 않으며, 최대 한도로 bit rate를 줄이는데 있다. 데이타 압축에 기여한 Shannon의 rate distortion 이론[7]에 의하면 스칼라 대신에 벡터를 코딩함으로써 더 좋은 성능을 얻을 수 있다는 것이다. 그러므로, 음성인식에 있어서 데이타 압축이라는 측면에서 VQ를 이용하였다. 즉, VQ는 입력 음성의 특징 벡터를 이미 저장되어 있는 특징 벡터들 중의 하나로 mapping시켜 주는 것을 의미한다.

이에 따라, 본 연구에서도 이 VQ codebook을 신경 회로망(MLP)에 이용하여 음성 인식을 행하고 VQ/HMM에 의한 음성 인식 방법과 비교하겠다. 학습 벡터들에 의해 codebook이 만들어지며, 입력 벡터는 codebook의 벡터들 중에서 최소의 거리를 갖는 벡터로 양자화 된다.

본 연구에서는 전 단어를 training data로 삼아 하나의 codebook을 취하는 방법을 사용하였다. 이 때 codeword를 구하는 방법은 clustering 기법 중 K-means 알고리즘[8]을 사용하였으며, 중심점 잡은 방법으로는 averaging 기법을 이용하였고 단모음 8개를 대상으로 인식 실험을 행한 관계로 codebook의 크기는 32로 하였다.

VQ에서 test set vector를 C_i라 하고 codebook entry를 C_m이라 했을 때, 국부적인 거리는 LPC cepstrum의 계수를 10차로 하였을 경우에 다음과 같다.[1]

$$d(C_m, C_i) = 0.04 * (C_{m0} - C_{i0})^2 + \sum_{p=1}^{10} (C_{mp} - C_{ip})^2 \quad (1)$$

2. VQ/ HMM 음성 인식[3]

HMM은 천이들에 의해 서로 연결된 상태들의 모임으로서 각 천이에는 2가지 종류의 확률이 관련되어 있다. 하나는 현재의 천이가 이루어질 천이 확률이고, 다른 하나는 천이가 이루어졌을 때 유한개의 관측 대상으로부터 각 출력 심벌이 방출되는 조건부 확률을 규정하는 출력 확률 밀도 함수(pdf)이다. 이 HMM에 의한 음성 인식에서의 모델의 파라메타 최적화를 위해서는 Baum-Welch의 reestimation 알고리즘을 이용하였으며, 인식 알고리즘으로는 Viterbi 알고리즘에 의한 인식보다 forward 알고리즘에 의한 인식이 더 좋다[4]는 고찰에 따라 forward 알고리즘을 사용하였다.

3. 제안된 VQ와 MLP에 의한 음성 인식

신경 회로망은 인간의 두뇌의 생물학적 신경 계통에 근거한 간단하고 많은 처리 요소들을 병렬로 상호 연결하여 학습을 통해 입력 패턴에 내재해 있는 정보를 분산 및 병렬 처리하는 정보 처리시스템으로, 기존의 음성 인식 알고리즘에서 해결할 수 없었던 패턴 인식 문제를 인공 신경 회로망을 이용하여 해결하려 한다.

본 연구에서 사용된 신경 회로망은 back propa-

gation 학습 알고리즘을 사용하는 MLP 구조로 이 구조의 입력 갯수는 VQ codebook의 크기와 같으며, 입력값은 VQ/HMM에서 와는 달리 codebook의 각 codeword가 어떤 데이타에 대해서 VQ의 열로서 선택될 확률값이다.

3.1 VQ를 이용한 입력값 결정

본 연구의 neural net의 입력 특징 파라메타는 VQ codebook[11]을 이용하여 구하게 되는데, 앞에서도 언급한 바와 같이 신경 회로망으로는 multi-layer perceptron(MLP)을 이용하는데 이때 입력 특징 파라메타의 수는 codeword의 수와 같으며 따라서 codebook의 크기를 32로 하여 실험한 관계로 본 연구에서는 32개의 입력값을 가지게 된다. 여기서 입력값을 구하는 방법은 C_m의 집합을 이용한다. 즉, m번째 입력값을 x_m이라 했을 때 다음 식에 의해 각 codeword에 의한 입력값이 결정된다.

$$x_m = \frac{1}{N} \sum_{n=1}^{N} \delta(C_n, C_m) \qquad (2)$$

여기서,

$$\delta(C_n, C_m) = \begin{cases} 1 & \text{if } C_m = C_n \\ 0 & \text{if } C_m = C_n \end{cases}$$

이고, N은 test data의 frame 수이다.

3.2 학습 및 인식 방법

패턴 인식에 널리 쓰이는 multi-layered perceptron의 구조는 그림 1에서 보여주는 바와 같이 hidden layer를 포함하는 feedforward multi-layered network 구조를 갖는다. 본 연구에서는 사용한 학습 알고리즘은 generalized Delta rule에 근거한 Rumelhart의 back propagation 알고리즘이다. 이 알고리즘에 의한 학습 절차는 크게 두 단계로 나눌 수 있는데, 그 첫단계는 인공 신경 회로망에 입력값을 제시하고 각 노드에 대해서 network의 입력 함수와 활성화 함수를 이용해서 출력을 산출하는 forward pass이고, 두번째 단계는 desired output과 actual output과의 차이를 계산하여 이 차이를 back propagation 시키면서 layer와 layer 사이의 가중치를 조절하는 backward pass이다. 이 두 단계는 시스템이 안정 될 때까지 즉, total error sum이 error criterion을 넘지 않을 때까지 계속 반복 실행한다.

따라서, 이런 개념에서 본 논문에서 이용한 back propagation 학습 알고리즘은 참고 논문 [9]에 따른다.

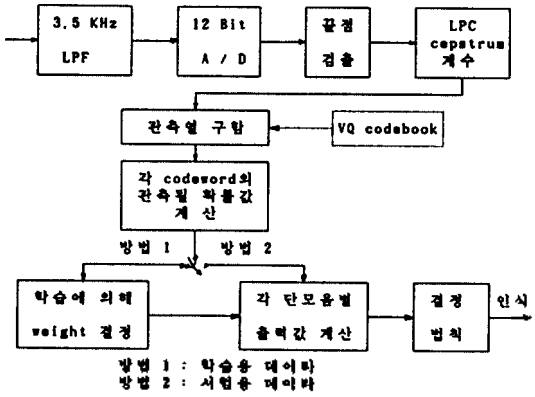

그림 1. MLP의 구조
Fig 1. Architecture of MLP

3.3 실험 조건

본 연구에서는 단모음 중 이중 모음화한 "ㅟ"와 "ㅢ"를 제외한 8개의 단모음을 대상으로 단모음 인식을 행하였는데 이중 "ㅔ"와 "ㅐ"를 구별해서 발음하는 화자들이 없는 관계로 "ㅔ"와 "ㅐ"는 같은 발음으로 간주해서 학습시에도 같은 class로 생각하고 학습시켰으며, 인식시에도 구별없이 같은 발음으로 간주하여 인식시켰다.

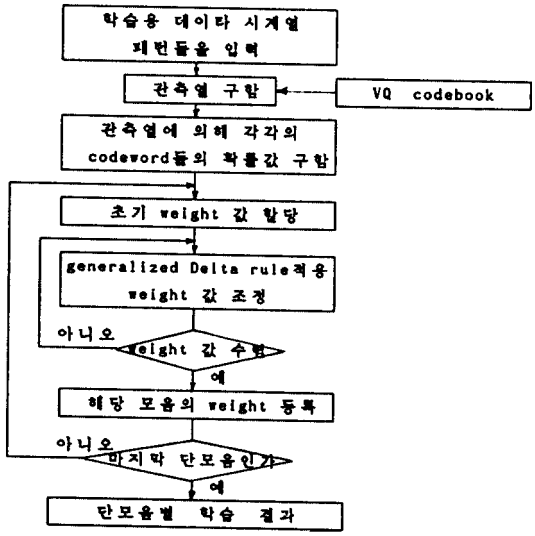

그림 2. VQ와 MLP의 학습 방법
Fig 2. The training method of VQ and MLP.

먼저, 인식 8개의 단모음을 "ㅐ"와 "ㅔ"를 같은 class로 간주하여 7개의 출력 node를 주어 학습 시켜 보았는데, 이 경우에 있어서는 1000번 정도까지 반복하여 학습시킨 후에는 학습시의 에러가 20000번까지 반복해서 학습시켜도 개선되지 않았다. 이때의 에러 값은 0.05에서 0.09까지 진동하였으며 그 이상의 개선이 되지 않았다. 또한 이 방법에 의해 인식시켰을 경우도 대부분이 인식되지 않았다.

그 이유는 back propagation 알고리즘은 정적 특징을 학습시키는데는 적합하나 음성은 정적 특징과 함께 동적 특징도 있어 그 출력 node 수를 증가시킬 경우에도 특징간의 classification을 잡아내지 못하기 때문이다.

따라서, 본 연구에서는 MLP의 병렬성은 고려하지 않고 HMM처럼 학습 능력만을 고려하여 실험하였다. 즉 8개의 단모음을 7개의 class로 간주하여 7개의 출력값을 고려한 것이 아니고, 학습 데이타를 가지고 각 단모음별로 그 단모음에 속하는지 않는지를 고려하여 weight 값을 조절해 줌으로써 각 단모음별로 학습시켰다. 이 때 sigmoid 함수에 사용된 desired output은 class에 속할 때 0.9이고 class에 속하지 않을 때 0.1을 주었다. 이에 대한 플로챠트는 그림 2과 같다.

위에서 말한 것처럼 학습된 데이타는 입력 데이타가 들어 오면 각 단모음별로 MLP를 통과하여 그 결과값을 조사하여 가장 값이 큰 것으로 나타난 단모음이 인식된 것으로 간주하였다. 이에 대한 플로챠트는 그림 3와 같다.

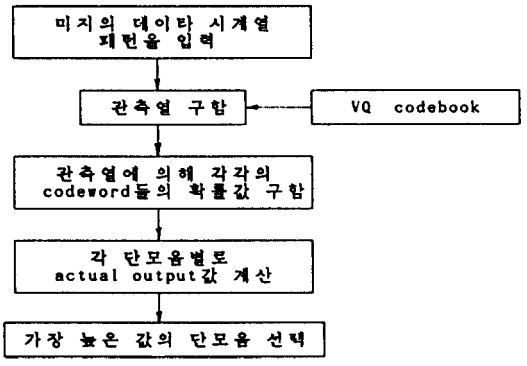

그림 3. VQ와 MLP의 인식 방법
Fig 3. The recognition method of VQ and MLP.

여기서, hidden layer의 수는 2이고, 입력 layer의 노드 수는 32이며, 1차 hidden layer의 노드수는 10을 주었으며, 2차 hidden layer의 노드수도 마찬가지로 10을 주었다. 또한 gain(learning rate)은 0.9로 하였고, momentum은 0.7을 주었다[10].

Ⅲ. 실험 및 결과

본 연구는 VQ와 MLP에 의한 인공 신경 회로망을 이용한 음성 인식 방법으로써 실험에 사용된 데이타는 단모음 8자 /ㅏ/, /ㅣ/, /ㅜ/, /ㅔ/, /ㅗ/, /ㅡ/, /ㅓ/, /ㅐ/로써 10명의 화자가 10번 발음한 총 800개의 음성으로써 이 중 세 화자가 각 단모음에 대해 5번씩 발음한 데이타를 가지고 학습을 시켰으며, 나머지 데이타를 가지고 실험하였다. 실험에서 세 화자의 데이타를 가지고 학습시켜 본 연구와 같은 방법이 얼마나 학습 능력이 좋은가를 실험해 보기 위해서이다. 따라서 본 연구에서 제안하는 방법 이 외에도 비교를 위해서 VQ/HMM에 의한 방법도 실험하였다.

1. 인식 시스템 구성

본 논문에서 제안하는 VQ와 MLP에 의한 음성 인식 시스템은 그림 4와 같다. 데이타는 연구실에서 마이크를 사용하여 IBM-386에서 받았는데 3.5KHz의 LPF를 통과시킨 후 8KHz의 샘플링 주파수하에서 A/D 변환기를 통해 12비트로 양자화된 데이타이다. 이렇게 받아진 데이타는 시작점과 끝점을 검출한 후 특징 벡터로 LPC cepstrum 계수를 구하고 이것을 VQ codebook의 codeword와 거리를 비교하여 관측열을 구한 후, 각 codeword가 가질 수 있는 확률을 계산해 이 codeword가 가질 수 있는 확률을 가지고 MLP의 입력으로 삼아 학습 및 인식을 시켰다.

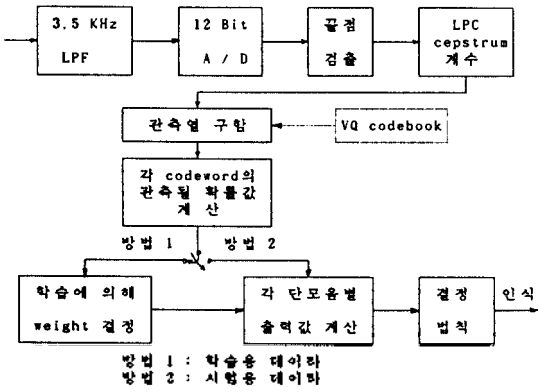

그림 4. VQ와 MLP 인식 시스템
Fig 4. Recognition system of VQ and MLP

2. 인식 결과

2.1 VQ/ HMM에 의한 인식 결과

실험에서 처음 세 화자가 각 단모음에 대해 5번 발음한 것으로 학습시켰으며 그 나머지 발음으로 인식 실험을 행했다. 이 때 본 논문은 단모음에만 적용시킨 관계로 state수는 3으로 하였으며, 관측 심볼수(codebook size)는 32로하여 실험하였다. 인식 실험에서 나타난 에러의 갯수를 표 1에 나타내었다. 실험 결과를 살펴 보면, 학습시에 사용한 A, B 및 C 화자의 5번씩 발음한 단모음의 인식은 100%였고, B 및 C 화자의 나머지 발음도 1개씩의 에러를 나타내었으며, 총 인식률은 91.63%이다.

표 1. VQ/HMM에 의한 에러의 갯수
Table 1. The number of errors by VQ/HMM

(단위 : 갯수)

단모음 / 화자	ㅏ	ㅣ	ㅜ	ㅔ	ㅗ	ㅡ	ㅓ	ㅐ	전체
A 화자	0	0	0	0	0	0	0	0	0
B 화자	0	0	0	0	0	1	0	0	1
C 화자	0	0	0	0	0	0	1	0	1
D 화자	1	3	0	1	4	6	0	0	15
E 화자	0	0	1	0	2	3	1	0	7
F 화자	0	0	7	0	2	1	1	0	11
G 화자	0	1	2	1	0	3	0	0	7
H 화자	0	4	0	2	3	1	0	0	10
I 화자	0	0	2	0	0	1	1	0	4
J 화자	2	0	4	0	3	1	1	0	11
전 체	3	8	16	4	14	17	5	0	67

2.2 제안된 VQ와 MLP에 의한 인식 결과

본 연구에서 제안하는 실험에서도 VQ/HMM에서와 마찬가지로 처음 세 화자가 각 단모음에 대해 5번씩 발음한 것으로 학습시켰으며 그 나머지 발음으로 인식 실험을 행했다. 이 때 입력 노드수는 각 단모음별로 관측 심볼수(codebook size)와 마찬가지로 32이며 이들 노드의 값은 관측 심볼들이 관측될 확률값이다. 인식 실험에서 나타난 에러의 갯수를 표 2에 나타내었다. 실험 결과를 살펴 보면, 학습시에 사용한 A, B 및 C 화자의 5번씩 발음한 단모음의 인식은 100%였고, B 및 C 화자의 나머지 발음도 1개씩의 에러를 나타내었으며, 총 인식률은 94.63%이다.

표 2. VQ와 MLP에 의한 인식률
Table 2. The number of errors by VQ and MLP

(단위 : 갯수)

단모음 / 화자	ㅏ	ㅣ	ㅜ	ㅔ	ㅗ	ㅡ	ㅓ	ㅐ	전체
A 화자	0	0	0	0	0	0	0	0	0
B 화자	0	0	0	0	0	1	0	0	1
C 화자	0	0	0	0	0	0	1	0	1
D 화자	0	1	0	0	2	3	0	0	6
E 화자	0	0	0	0	2	3	1	0	6
F 화자	0	0	3	0	0	1	0	0	4
G 화자	0	0	2	0	0	2	0	0	4
H 화자	0	2	0	2	3	1	0	0	8
I 화자	0	0	2	0	0	2	1	0	5
J 화자	1	0	5	0	2	0	0	0	8
전 체	2	3	12	2	9	12	3	0	43

Ⅳ. 결 론

본 연구는 VQ와 MLP에 의한 음성 인식에 관한 논문으로 VQ codebook으로 부터 관측열을 구하고 이를 토대로하여 각 codeword가 데이타로부터 가질 수 있는 확률값을 계산하여 그 값을 입력값으로하여 MLP에 의해 인식하는 시스템을 제안하였다.

본 연구에서 제안된 방법은 몇가지 방식에서 VQ/HMM에 의한 실험과 비슷한데 첫째 VQ codebook을 이용하여 관측열을 구한다는 점이고, 학습시에 개별적인 class 단위로 이루어진다는 점이다.

또한 본 연구의 목적은 같은 VQ codeword하에서 HMM에 의해 학습하여 인식 실험한 경우와 MLP에 의해 학습하여 인식 실험한 경우가 같은 조건하에서 제안된 방법이 기존의 방법에 비해 얼마나 정확하게 분류해 내는지에 대해 알아보고 앞으로 인식 알고리즘으로써의 유용성을 알아보는데 있다.

실험에 의하면 학습된 데이타에서는 모두 100% 인식률을 나타내었고, 같은 화자의 학습되지 않은 데이타에서도 모두 1개의 에러로 99.2%의 높은 인식률을 나타내었다. 그러나, 학습에 참여하지 않은 화자의 인식에 있어서는 차이를 나타내어, VQ/HMM의 경우는 89.00%의 인식률을 나타내었고, VQ와 MLP에 의한 인식의 경우는 92.68%의 인식률을 나타내었다.

따라서, 실험 결과에 의하면, 본 연구에서 제안하는 VQ와 MLP에 의한 인식 실험이 기존의 VQ/

HMM에 의한 인식 방법보다 학습 능력이 우수함을 알 수 있었다.

그러나, 앞으로 더 연구되어야 할 과제는 시간적 변화를 어떻게 좀 더 효율적으로 모델링할 것인가 하는 문제인데, 이것은 MSVQ 방법등을 이용하여 극복해야 할 것으로 보이며, 제안된 방법을 MSVQ 방법이나, Level Building 방법에 적용하여 이용하면 단독어 인식, 연결어 및 연속음 인식에도 사용될 수 있으리라 본다.

참 고 문 헌

1. 김순협, "한국어 음성의 분석과 자동 인식에 관한 연구," 박사 논문, 연세대학교 대학원, 1982. 12.
2. Y. Linde, A. Buzo, and R.M. Gray "An algorithm of Vector Quantizer Design," IEEE Trans. Comman, Vol.COM-28 pp.84-95, Jan 1980.
3. L,R,Rabiner, B.H.Juang, "An Intorduction to Hidden Markov Models," IEEE ASSP MAGAZINE JAN. 1986.
4. 안태옥 외, "DHMM을 이용한 한국어 음성 인식," 한국음향 학회, 제10권 1호, pp.52-60, 1991. 2.
5. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme Recognition using Time-Delay Neural Network," IEEE Tran. of Acoustics, Speech and Signal Processing, Vol. 1, ASSP-37, March 1989.
6. T. Kohonen, G. Barna and R. Chrisley, "Statistical Pattern Recognition with Neural networks : Benchmarking Studies," IEEE, Proc. of ICNN, Vol.1, pp. 61-68, July 1988.
7. C. E. Shannon, "A mathematical Theory of Communication," Bell Sys. Tech. J. 27, pp.379~423, 623~656, 1948.
8. J.T. Tou, R.C.Gonzalez, Pattern recognition Principles, Addison-Wesley Publishing Company, Inc, 1974.
9. R. P. Lippmann, "An Introduction to Computing with Neural Nets," IEEE ASSP Magazine, 4-22, April 1987.
10. Yuh-Han Pao, Adaptive Pattern Recognition and Neural Networks, Addison-Wesley Publishing Company, 1989.
11. 이성권, "VQ를 이용한 DDD 지역명 인식에 관한 연구," 석사 학위 논문, 광운대학교 대학원, 1989. 12.

▲안태옥 : 10권 1호 참조

▲김순협 : 10권 1호 참조

▲이 상 훈(Sang Hoon Lee) 1958년 8월 2일생

978년 3월~1983년 2월 : 광운대학교 응용전자공학과 전자공학사

1983년 3월~1987년 8월 : 광운대학교 대학원 전자과전자공학석사

1988년 3월~1992년 2월 : 광운대학교 대학원 전자과 전자 공학박사

1990년 9월~현재 : 광운대학교 전자계산교육원 전임강사