特輯

技術解說

신경회로망을 이용한 음성인식의 연구동향

— Study Trends of Speech Recognition using Neural Networks —

이 기 영* · 배 철 수* · 최 갑 석**

(*관동대학교 전자통신공학과, **명지대학교 정보통신공학과)

Ⅰ. 서 론

아직까지 자연스러운 음성을 맨-머쉰 인터페이스에 사용하고 있지 못한 주된 이유로는, 음성신호의 변질성이나 중첩성, 대어휘(5만단어 내외)의 실시간 처리, 다양한 해석(음성학, 음운학, 구문론, 의미론 등)의 종합성 및 음성인식의 포괄적이론의 결핍 등을 들 수 있다[1]. 현재의 맨-머쉰 인터페이스를 위한 연구에서는 인식대상을 학습단어, 고립단어, 소어휘, 특정화자 및 제한된 문법범위로 국한시켜 비교적 높은 인식결과를 얻고 있으나, 자연스러운 음성에 의해 맨-머쉰 인터페이스가 이루어지기 위해서 기계로 하여금 그 능력을 갖도록하기 위한 대어휘-연속음성-불특성정화자-실시간 처리라는 목표아래 새로운 모델이나 기법의 연구가 계속되고 있다.

최근들어 음성인식의 최고성능을 구가하고 있는 모델인 *HMM*[2-4]은, Viterbi 알고리즘의 효과적 디코딩방법과 forward-backward 알고리즘의 자동지도학습방법을 지닌 음성의 디지탈신호처리에 매우 적당한 인식모델로 알려져 있다. 그러나, HMM의 한계영역이 있다면 그자신의 모델에서 음향/음운성 모델링에 약하여 유사음성에 대한 처리가 곤란하며, 유한상태나 확율적 문법만을 받아 돌이는 모델이기 때문에 이해 수준이나 semantic 모델링에로의 적용이 곤란하다. 이에 대해 인간의 사고영역으로 접근하려는 모델인 신경회로망은 음성인식에 필요한 전처리나 거리계산 및 비선형 시간 정규화 등의 문제를 흡수할 수 있는 잠재적인 능력 때문에 HMM 모델의 대체적인 모델로서 신경회로망을 이용한 음성인식의 연구에 박차가 가해지고 있으며, 최근에는 음성인식에 필요한 여러 과제들을 기존의 모델보다 효과적으로 해결하기 위하여 이들을 이용한 음성인식의 연구에 많은 관심이 집중되고 있다[5-28].

신경회로망을 이용한 음성인식의 연구는 아직까지 불특정화자 대어휘음성인식에 미치지 못한 현수준에서 새로운 접근방식으로 각광을 받고 있으며, connectionism 또는 parallel distributed processing 이라고도 불리우고 있다. 본고에서는 신경회로망을 이용한 음성인식기법 또는 모델에 대하여 소개하고자 하며, 음성인식을 위한 접근방법에 따라 정적인 신경회로망과 동적인 신경회로망으로 나누어 여러 형태의 신경회로망에 관하여 최근까지의 연구 동향을 서술하였고, 마지막으로 장래의 과제에 대하여 결론을 맺었다.

Ⅱ. 정적인 신경회로망

정적인 신경회로망을 이용하여 음성을 인식하기 위해서는 단어 및 모음이나 각 음소에 따라 추출된 스펙트럼특징으로 정적인 입력패턴을 구성하여 신경

회로망에 입력시키는 것이 일반적이다. 본절에서는 이연구범위를 다층퍼셉트론와 계층신경회로망의 두 가지로 나누어 기술하고자 한다.

2-1. 다층퍼셉트론(multilayer perceptrons)

정적인 모델에서 신경회로망의 입력은 주어진 음성신호를 대표할 수 있는 정적패턴 또는 특징량으로 하였으며, 출력에 의해 분류가 수행된다. Huang 등[7]은 포만트주파수{F_1, F_2}를 입력으로한 3층 역전파신경회로망에 의해 모음음성의 비선형분류영역을 구성하였으며, 그림 1에 나타내었다. 이로부터 신경회로망이 비선형분류기로써의 기능이 있음을 확인하였다.

Elman등[8]은 모음 /a, i, u/에 뒤따르는 유성파열음 /b, d, g/의 음운분류실험을 보고하였으며, 학습데이타에서 잡음을 제거하였을 때 더 향상된 인식율을 얻으므로써 잡음이 학습데이타의 특이성을 잃게 할 수 있다는 사실을 확인하였다. 또한, Lippmann등[5]은 고립숫자음인식에서 각 숫자음의 최대에너지구간의 특징벡터를 신경회로망의 입력으로 하여 인식을 시도하였으며, Peeling등[9]은 고립 숫자음의 인식에서 1.2초 동안의 특징벡터를 신경회로망의 입력으로 할 때 단어의 발성이 짧으면 그 나머지 부분을 '0'으로 대체한 결과 높은 인식율을 얻었다. 국내에서도 한국어 음성인식을 위한 다층퍼셉트론에 관한 연구[10-12]가 활발히 진행되고 있다. 표1에는 이들의 연구에서 사용한 네트워크의 구조, 음성데이타 및 고찰사항에 관하여 정리하였다.

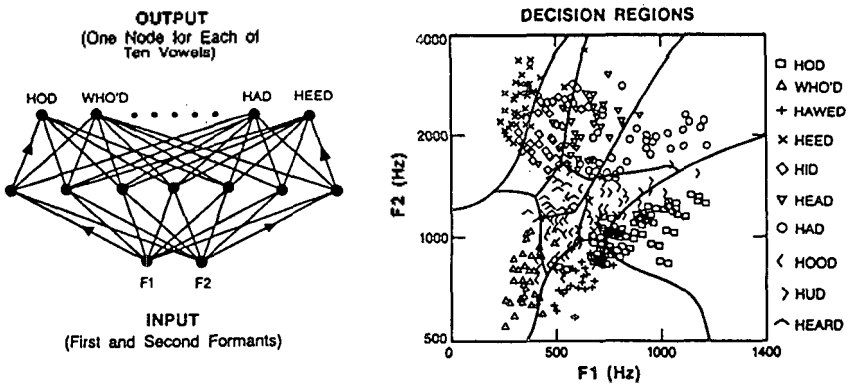


그림. 1. 모음의 포만트{F_1, F_2}를 이용한 3층 역전파 신경회로망(MLP)과 모음분류의 경계
Fig 1. A three-layer backpropagation network(NLP) used to form classification boundaries on the formants F_1, F_2 for vowels

표 1. 정적인 신경회로망을 이용한 음성인식
Table 1. Speech recognition using static neural network

Study	Network	Speech material	Remarks
Huang, Lippmann	Input : F1, F2 Hidden : 50 Output : 10	10 Vowels	• 50,000 Trials • Nonlinear Classifier
Elman, Zipser	Input : 16 BPF x 20 Frames Hidden : 16-256 Output : 20	9CV Syllables : /b, d, g/x /i, a, u/	• Segmented manually • Small Data Base for Training

Lippmann, Gold	SLP, MLP Input : 2x11 Cep. Output : 20	TI-20 Word Data Base 2 Maxmum-energy Frames for Each Word	• Outperformed a Gaussian Classifier but not kNN Classifier • SLP sometimes never converged
Kammerer, Kupper	SLP Input : 16 Frames x 16 bit-spectrum	TI-20 Word Data Base Time-normalized to 16 Frames	• Speaker DEP. : slightly better than DTW • Speaker IND. : poorer than DTW
Peeling, Moore	MLP Input : 19CH. BPF x 60 Frames (1.2sce)	RSRE 40-speaker Digit Data Base 1.2sec be enough for the longst word. Shorter word were padded with zero.	Error Rates Talker DEP. : 0.25% Multi-talker : 1.9%
Keun Sung Bae	MLP 10 x 3 Segments x 5, 3, 1 Frames : Inputs	Korean Digits(11) 5 Talkers 1100 Tokens Features : Cep., ∂Cep., Reflect	• 3 Segments are efficient • (Cep. +∂Cep.) is better than one feature
Haingsei Lee	Composite NN which consists of contrl net and sub-nets	3 Talkers Korean Consonants with /i/ 100 Tokens	• The control net identifies the group. and sub-nets recognize the input pattern in each group.

2-2. 계층신경망(hierarchical neural nets)

본절의 계층신경망은 음성패턴을 분류하기 위하여 커널함수(kernel function)를 계산해 낸다. 이 신경망은 학습시간이 빠르며 지도 및 비지도 학습데이타를 조합하여 사용하는 특징이 있다.

Huang과 Lippmann[13]은 Kohonen의 feature-map[14, 15]을 다층퍼셉트론에 의해 분류했던 그림 1의 모음에 적용하여 그 성능을 평가한 바 있으며 그림 2에 그 블록도를 나타내었다. 이 신경망에서 중간코드북(intermediate codebook)에 대한 node는 입력과 집단중심의 유클리디안거리에 관계된 커널함수를 계산해 내도록 되어 있다. 이 그림에서 하단층에서는 벡터양자화를 구성하기 위하여 먼저 비지도적으로 학습하고 상단층에서는 변형된 LMS 알고리즘을 이용하여 지도적으로 학습을 수행한다.

Kohonen[14, 15]등이 발표한 LVQ(learning vector quantization)는 feature-map 그 구조는 유사하지만

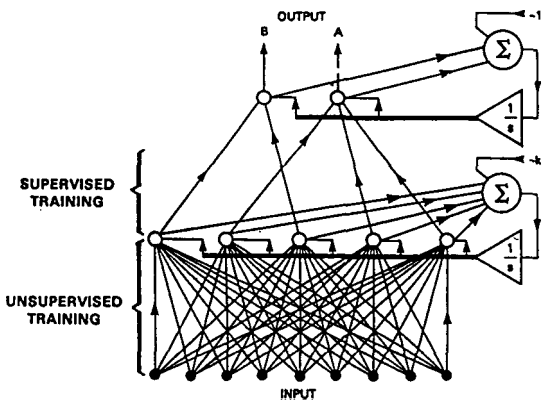


그림 2. 계층 feature-map 분류기의 블록도
Fig 2. Block diagram of hierarchical feature-map classifier

중간코드북의 가중치를 조절하기 위한 지도학습과정이 서로 다르다. 컴퓨터 Masscomp MC5600을 이용

하여 학습을 수행한 결과 볼쯔만머쉰은 5시간, 다층 신경망은 1시간, 그리고 LVQ는 20분 정도의 학습시간이 소요되어 다층신경망과 유사한 성능을 가지면서 학습시간이 고속화된 신경망으로도 알려져 있다. 표2에는 이들의 연구에서 사용한 네트워크의 구조, 음성데이타 및 고찰사항에 관하여 정리하였다.

표 2. 계층신경망을 이용한 음성인식
Table 2. Speech recognition using hierarchical neural network

Study	Network	Speech material	Remarks
Huang, Lippmann	MLP, Feature Map Classifier (FMC) 2 Inputs	67 Talkers 10 Vowels 671 Tokens	• FMC trains faster than MLP
Kohonen	LVQ simillar to FMC	1550 Single – Frame Patterns Manually Ext – racted	• Slightly better than kNN, Baysian in recognition. • Training is faster than MLP, Boltzman.

Ⅲ. 동적인 신경회로망

정적인 신경회로망을 이요한 음성인식으로 술單語 또는 단어 단위보다 작은 단위의 특징벡터의 행렬릉 입력으로 하고 출력의 각 유니트에 의해 단어를 분류하기 때문에 어휘수를 증가시키면 신경회로망의 크기나 그에 따른 학습시간이 증가하는 문제가 생기며, 이 신경망의 입력패턴은 주파수성분과 시간성분을 모두 포함하여 구성되기 때문에 세그멘테이션시간과 특징추출을 위한 처리시간을 소요하므로 실시간처리에도 부적합하다. 실제의 연구에 있어서 세그멘테이션의 경우 그 정확도가 신경망에 의한 인식성능에 지대한 영향을 미치기 때문에 대부분 수동으로 세그멘테이션하는 경우가 많다.

이에 대하여 단시간지연, 시간축의 집적화 및 귀환연결(recurrent connections) 등으로 구성되는 동적인 신경망은 음성인식을 위하여 특별히 개발되었으며, 이신경망의 입력은 순차적으로 프레임단위를 대상으로 하기 때문에 프리세그멘테이션을 필요로 하지 않으므로 실시간처리에 매우 유용한 신경망모델이다. 따라서, 신경회로망으로 하여금 학습데이타의 세그멘트 경계에 의존치 않고 음성신호의 특징을 파악할 수 있는 시불변성의 확보방법으로 동적인 접근방식의 신경회로망 모델의 연구가 시작되었다. 다층퍼셉트론의 출력노드에 시간축 지연 및 시간축 집적형태를 적용하여 좋은 결과를 얻은 Waibel의 TDNN이나 Watrous의 recurrent 형태의 다층퍼셉트론 등은 동적인 접근방식의 좋은 예이다. 본절에서는 동적인 신경망을 몇가지로 나누어 기술하였다.

3-1. Time Delay Neural Network(TDNN)

TDNN은 은둔층의 유니트들이 시간지연 입력벡터들과 연결되어 있으며, 입력벡터로부터 활성패턴을 생성하고 그의 상층에서는 시간위치와 관계없이 시불변적으로 활성패턴에 존재하는 음운의 존재 여부를 파악한다. 그림 3에 TDNN의 기본 블록도를 나타내었으며 표3에는 TDNN을 이용한 동적인 신경회로망의 연구들을 정리하였다. 여기서 Waible[6]이 TDNN을 이용하여 수작업의 세그멘테이션을 통한 98.5%의 인식율을 얻었으며 동일 데이타에 대하여 이산HMM을 이용한 결과 93.5%의 인식율을 얻어 TDNN의 우수성을 확인하였다. 또한, 1988년 Lang과 Hinton[7]은 TDNN을 이용하여 "B", "D", "E", "V"에 대한 인식실험을 하였으며, 학습시간을 반복시키는 방법도 연구되었다. 이연구에서 다수 화자에 대한 오인식율이 7.8%로 비교적 크게 나타나고 있다. 국내에서도 한국어 음성인식을 위한 TDNN에 관한 연구[20]가 활발히 진행되고 있다.

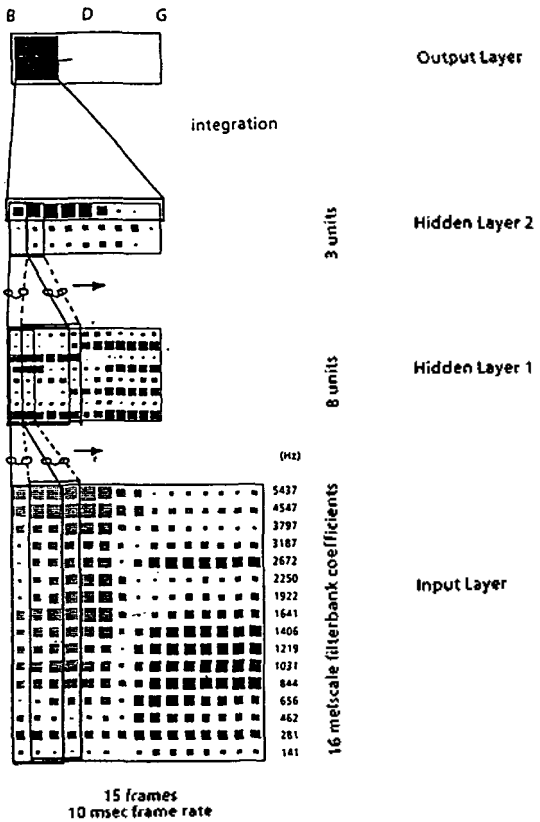


그림 3. 파열자음 /b, d, g/에 대한 TDNN
Fig 3. TDNN for stop consonants /b, d, g/

3-2. LVQ & Recurrent Nets

McDermott과 Katagiri[21]등은 Waibel이 TDNN에서 사용한 바 있는 동일한 음성데이타 /b, d, g/에 Kohonen의 LVQ와 그림 3의 시간지연망의 구조와 결합하여 분류를 수행하여 Waibel의 연구와 비슷한 결과를 얻었다. 이신경망에서는 커널함수의 계산과 시간지연망을 위한 동작을 모두 수행하기 때문에 많은 계산량을 요구하며, 가중치의 수가 TDNN에 비해 약 30배 가량이 많기 때문에 더 큰 기억용량을 요구한다. 그러나 LQ가 정적인 신경망으로 사용될 경우와 같이 학습시간은 다른 시경망에 비해 고속인 장점이 있다. 그림 4에 동적인 신경망으로 사용되는 LVQ의 블록도를 나타내었다.

Recurrent nets의 연구도 시불변성의 음운인식을 위해 1988년 Watrous[22]에 의해 개발되었다. 처음에는 학습방법, 해석 및 설계의 어려움으로 음성인식분야에 적용되지 않았으나, recurrent Bolzmann machine[23]을 기점으로 연구가 시작되었으며, 과거출력을 현재의 context 층으로 피드백하므로써 서로 상관성을 갖도록 한다는 것을 목적으로 연구되고 있으며 학습시간이 긴 단점의 극복과 시불변적 인식을 위해 변경된 역전파학습알고리즘의 연구등이 계속되고 있다. 그림 5에 recurrent net의 기본 블록도를 나타내었다. 국내에서도 이를 위한 한국어 음성인식의 연구

표 3. 시간지연신경망을 이용한 음성인식
Table 3. Speech recognition using TDNN

Study	Network	Speech material	Remarks
Lang, Hinton	Tine Delay MLP 16 Inputs	100 Talkers "B, D, E, V" 768 Tokens	• Some form of time integration in output nodes
Unnikroshnan, Hopfied, Tank	Time Concentration Net 32 Inputs	1 Talker Digits 432 Tokens	• Using Variable length delay • No better than HMM
Waibel et al.	Time Delay MLP 16 Inputs	1 Japanese Talker, /b, d, g/ 4,000 Tokens	• 15 Frames segmented by hand around voice onset • Better than HMM
Soon Hyob Kim	TDNN 16 mel-scaled spectral inputs	2 Talkers /b, d, g, t, p, k/ x /a, i, u, e, o/ 150 Tokens	• New training method with Cosh algorithm is better & faster than the error back − propagation method.

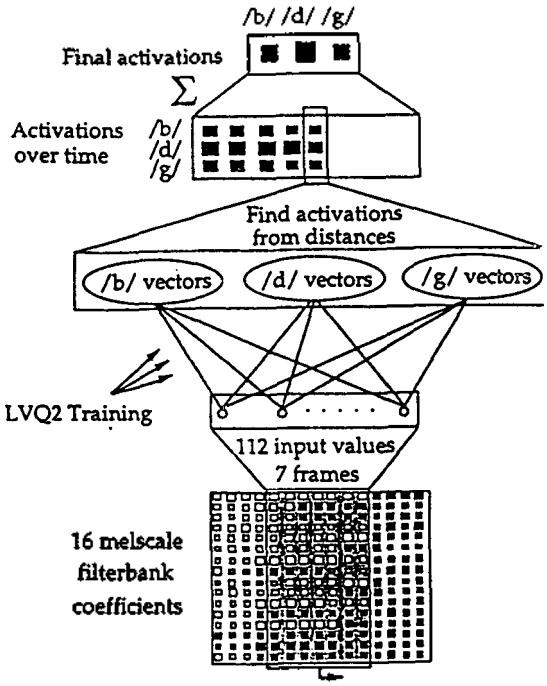


그림 4. 파열자음 /b, d, g/에 대한 LVQ
Fig 4. LCQ for stop consonants /b, d, g/

[16, 24, 25]가 활발히 진행되고 있다. 표4에는 이들의 연구에서 사용한 네트워크의 구조, 음성데이타 및 고찰사항에 관하여 정리하였다.

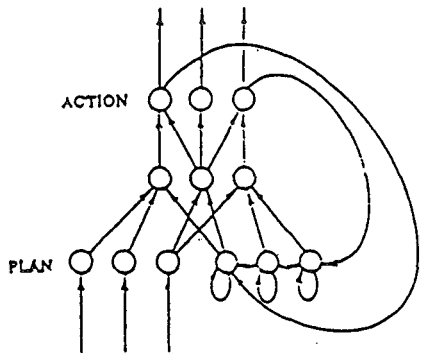


그림 5. Recurrent 망
Fig 5. Recurrent nets

3-3. 고전적인 방법과 신경망의 결합

음성인식의 성능향상이나 실시간처리의 효과성을 위해 신경회로망과 DTW나 HMM 같은 종래의 방법을 조합한 인식알고리즘들이 제안되기 시작하였다.

표 4. LVQ와 recurrent net 이용한 음성인식
Table 4. Speech recognition using LVQ and recurrent nets

Study	Network	Speech material	Remarks
McDermott, Katagiri	Time Delay LVQ 16 Inputs	3 Japanese Talkers, /b, d, g/ 4,000 Tokens	• Computation & Memory Capacity is more, but Training is faster
Hwang Soo Lee	LVQ2, MLVQ2 simllar to FMC	3 Talkers Balanced Words 330 Tokens Korean Phonemes	• MLVQ2 is slightly better than LVQ2.
Anderson, Merill, Port	Recurrent Net 36 Inputs	20 Talkers /b, d, g, p, t, k/ with /a/ 561 Tokens	• Best performance in recurrent nets
Robinson, Fallside	Recurrent Net 20 Inputs	7 Talkers 27 Phoneme 558 Sentences	• Trained with the modified form of backpropa — gation

Heeyeung Hwang	Recurrent Net 248 Inputs	1 Talker Vowels : 8 Plosives : 9	• Delta cepstrum is best in 6 features.
Kap Seok Choi	MLP with the Feedback Architecture	2 Talkers Korean Digits 100 Tokens	• Better than MLP without Feedback

예를 들면, DNN[26]은 DTW의 비선형시간축정규화 기능과 신경회로망의 학습 기능을 결합하여 단어인식 성능을 향상시키려는 목적으로 제안된 것이 있으며, HMM의 학습알고리즘과 신경회로망의 식별능력을 조합하여 제안된 높은 인식성능의 알고리즘[27]도 있다. 표5에는 신경회로망과 종래의 방법을 조합한 연구에 대하여 정리한 것이다. 그밖에 입력층의 구조와 시불변성의 확보번에서 과거의 특장벡터열을 신경회로망에 입력하여 현재의 특징벡터를 예측하도록 구성된 NPM[28, 29] 등이 있다.

Ⅳ. 장래의 과제

신경회로망은 음성인식을 위한 처리단계를 모두 흡수할 수 있다는 잠재적 능력을 가지고 있기 때문에 연구자들에 의해 많은 관심과 연구가 축적되어 오고 있다. 본논문에서는 음성인식을 위한 신경회로망의 최근까지의 연구 동향을 살펴보고 음성인식을 위한 접근방법에 따라 정적인 접근방식과 동적인 접근방식에 의한 신경회로망의 적용례를 기술하였다. 이상으로부터 신경회로망을 이용한 음성인식에서, 아직까지 인식성능이 높지는 못하지만, 우선 해결해야 할 문제를 살펴보면 다음과 같다.

1) 학습시 많은 학습데이타 양 및 학습시간의 감소.
2) 유사음절/음소의 불특정화자에 대한 음성인식의 개선.
3) 대어휘-연속음성의 인식.

이문제들을 해결하기 위하여, 부단어나 단어단위의 모델학습방법, 지도학습의 양을 줄이는 방법, 양질의 음운특성추출방법, 음향/음운성의 기본단위 분리방법, 비선형시불변성의 확보 및 학습시간의 감소방법 등의 연구가 필요하며, 많은 연구자들은 이를 위해 신경회로망과 고전적인 방법들을 조합한 방법에 의한 연구도 활발히 진행되고 있다.

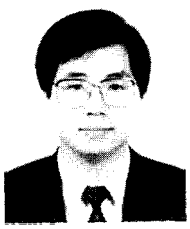
또한, 앞으로 긴 기간이 걸리더라도 점차적인 해결을 필요로하는 과제가 있다면 신경회로망으로 하여금, syntactic, semantic, pragmatic constraint를 학습할 수 있도록 하여 연속음성인식 및 이해 수준까지 신경회로망의 성능을 향상시키는 것이다.

참 고 문 헌

1. K.H.Klatt. "The Problem of Variablility In Speech Recognition and Models of Speech Perception," in Invatiance And Variablility in Speech Processes, eds. J.S.Perkell and D.H.Klatt, 300-324, N.J. : Lawrence Erlbaum, 1986
2. D.B.Paul, "A Speaker-Stress Resistant HMM Isolated Word Recognizer," in ICASSP 87, 713-716, 1987
3. A.Averbuch, L.Bahl, R.Bakis, "Experiments with Tangora 20,000 Word Speech Recognizer," in ICASSP 87, 701-704, 1987
4. Kai-Fu Lee, H.W.Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM," in ICASSP 88, 123-126, 1988
5. R.P.Lippmann, B.Gold, "Neural Net Classifiers Useful for Speech Recognition," Ist. International Conference on Neural Networks, IEEE, IV-417, Jun. 1987
6. G.E.Hinton, "Connectionist Learning Procedures," Technical Report CMU-CS-87-115, Carnegie Mellon Univ., Computer Science Department, 1987
7. W. Huang R.Lippmann, T.Nguyen, "Neural Nets for Speech Recognition," Conference of the Acoustic Society of America, Seattle WA, 1988
8. J.L.Elman, D.Ziper, "Learning the Hidden Structure of Speech, "Thecnical Report, Univ. California, San Diegi, Feb. 1987
9. S.Peeling, R.Moore, "Experiments in Isolated Digit Recognition Using the Muli-layer Perceptron, "Technical Report 4073, Royal Speech and Radar Establishment(RSRE), Dec.1987

10. 박현화, 강해동, 배건성, "신경망을 이용한 한국어숫자 음식에 관한 연구," 한국음향학회지 제11권 3호 1992
11. 김석동, 이행세, "신경망을 이용한 우리말 음성의 인식에 관한 연구," 한국음향학회지 제11권 3호 1992
12. 안태옥, 이상훈, 김순협, "VQ와 MLP를 이용한 단모음 인식에 관한 연구," 한국음향학회지 제12권 1호 1993
13. W.Huang, R.Lippmann, "Neural Net and Traditional Classifiers," Neural Information Processing System, ed. D.Anderson, 387-396, New York, America Institute of Physics, 1988
14. T.Kohonen, "Self-Organization and Associative Memory," Berlin, Springer-Verlag, 1984
15. T.Kohonen, "An Introduction to Neural Computing," Neural Networks 1, 3-16, 1988
16. 김홍국, 이황수, "수정된 LVQ2 알고리즘을 이용한 음소분류," 한국음향학회지 제12권 1E호 1993
17. A. Waibel, et al., "Phoneme Recognition Using Time-delay Neural Networks," IEEE Trans. Vol. ASSP-37, 328-339, Mar. 1989
18. K.J.Lang, G.E.Hinton, "The Development of the Time-Delay Neural Network Architecture for Speech Recognition," Technical Report CMU-CS-88-152, Carnegie Mellon Univ. 1988
19. J.Hampshire, A.Waibel, "A Novel Objective Function For Improved Phoneme Recognition Using Time-delay Neural Networks," IEEE Trans. Neural Networks, 1, 216-228, Jun. 1990
20. 최영배, 양진우, 이형준, 김순협, "한국어 음성인식을 위한 신경회로망에 관한 연구," 한국음향학회지 제13권 1호 1994
21. E.McDermott, S.Katagiri, "Phoneme Recognition Using Kohonen's Learning Vector Quantization," ATR Workshop on Neural Networks and Parallel Distributed Processing, Osaka, Japan, 1988
22. R.Watrous, "Speech Recognition Using Connectionist Networks," Ph.D. thesis, University of Pennsyvania, 1988
23. R.W.Prager, T.D.Harrison, F.Fallside, "Boltzmann Machines for Speech Recognition," Computer Speech and Language 1, 2-27, 1986
24. 김기석, 임은진, 황희융, "음성인식신경망을 위한 음성 파라미터들의 비교," 한국음향학회지 제11권 3호 1993
25. 이기영, 최정철, 최종환, 최갑석, "시퀀스-피이드백 신경회로망을 이용한 음성인식에 관한 연구," 한국음향학회지 제12권 5호 1993
26. H.Sakoe, et al., "Speaker-Independent Word Recognition Using Dynamic Programming Neural Networks," in ICASSP 89, 29-32, May 1989
27. G.Zavaliagkos, Y.Zhao, R.Schwartz, J.Makhoul, "A Hybrid Segmental Neural Net/Hidden Markov Model System for Continuous Speech Recognition," IEEE Transaction on Speech And Audio Processing, Vol.2, No.1, Jan. 1994
28. K.Iso, T.Watanabe, "Speaker-Independent Word Recognition Using a Neural Prediction Model," in ICASSP 90, 441-444, Apr. 1990
29. 이기영, 김한재, 김승겸, 최갑석, "NPU 선형매칭 한국어 단어 인식," 한국음향학회지 제11권 6호 1992

▲이 기 영 1961년 5월 7일생



1984년 2월 : 명지대학교 전자공학과 졸업
1986년 2월 : 명지대학교 대학원 전자공학과 석사과정 졸업(공학석사)
1992년 2월 : 명지대학교 대학원 전자공학과 박사과정 졸업(공학박사)
1993년 3월 ~ 현재 : 관동대학교 전자통신공학과 조교수
※관심분야 : 음성인식, 성질변환.

▲최 갑 석
제10권 4호 참조