Predictive RBFN을 이용한 단독 숫자음 인식

한학용*, 김상범**, 김주성*, 김수훈*, 허강인*

* 동아대학교 전자공학과, **섬유기능대학 컴퓨터공학과
kihur@seunghak.donga.ac.kr

Recognition of isolated digits using Predictive RBF Network

Hag-Yong Han*, Sang-Berm Kim**, Joo-Sung Kim*, Soo-Hoon Kim*, Kang-In Hur*

* Dept. of Electronic Engineering, Dong-A Univ.
** Dept. of Computer Science, Textile Polytechnic College

Abstract

본 논문에서 제안한 예측형 RBFN(Radial Basis Function Network)은 HMM과 신경망을 결합한 하이브리드 구조이다. 이 신경망은 HMM으로 추정한 확률분포 파라미터를 사용하여 중간층의 활성화 함수의 출력을 결정하고, 중간층과 출력층의 연결강도만 네트워크 내에서 학습한다. 그리고 HMM으로 추정한 확률분포 파라미터는 두 가지 방법으로 예측형 RBFN에 이용하였다. 첫 번째는 HMM의 각 상태의 혼합수 만큼의 중간층 유니트를 주는 방법이고, 두 번째는 HMM의 혼합수×출력분포수 만큼의 중간층 유니트를 주는 방법이다.

실험결과, 예측형 RBFN은 다른 방법들의 결과보다 4.5~6.5% 저하된 결과를 보였지만 다른 신경망에 비해서 학습 반복 횟수를 작게할 수 있었으며 전체 학습시간을 대폭 단축할 수 있었다.

Ⅰ. 서론

음성은 인간의 가장 자연스러운 의사소통의 수단이며 음성에 의한 인간-기계의 인터페이스는 속도가 빠르고 특별한 훈련 없이도 이루어진다. 또한 컴퓨터 및 정보통신 기술의 발전으로 음성인식 기술은 중요한 연구과제가 되고 있다.

음성인식 방법에서 HMM은 화자의 개인차등에 다른 음성패턴의 변동을 통계적으로 처리한 후 그 통계량을 확률적인 형태의 모델에 반영하여 인식하는 방법이다. 이 방법은 개인차나 조음 결합 등의 영향으로 나타나는 음성 패턴의 변동이 보다 정확하게 반영되고 확률 통계론에 의한 이론적 전개가 용이하며, 음소나 음절 단위의 모델을 단어·문장 등의 단위로 쉽게 확장할 수 있는 장점이 있다. 그러나 모델의 구조를 결정할 때 시행착오나 경험에 으존하는 경우가 많고, 학습시에는 다량의 샘플데이터와 계산능력이 필요하고 음성의 과도적인 정보를 경시하는 경향 때문에 패턴의 시간적 상관 표현력이 부족한 결점이 있다. [1][2]

그러나 신경망은 화자의 차이 등으로 나타나는 스펙트럼의 변동을 네트워크의 연결강도로 처리할 수 있고 한번에 많은 프레임의 데이터

를 입력할 수 있다. [3][4]

따라서 본 논문에서는 HMM과 신경망을 결합한 RBFN을 구성하였다. HMM으로 추정한 확률분포 파라미터를 사용하여 중간층의 활성화 함수의 출력을 결정하고, 중간층과 출력층의 연결강도만 네트워크 내에서 최급 하강법으로 학습한다. 따라서 MLP에 비해 학습시간을 대폭 단출할 수 있다. 그리고 HMM으로 추정한 확률분포 파라미터는 두 가지 방법으로 예측형 RBFN에 이용한다. 첫 번째는 HMM의 각 상태의 혼합수 만큼의 중간층 유니트를 주는 방법이고, 두 번째는 HMM의 혼합수×출력분포수 만큼의 중간층 유니트를 주는 방법이다.

따라서 본 논문에서 제안한 RBFN의 인식성능은 HMM, MLP, Jordan형 RNN, Elman형 RNN등의 방법과 비교하였다.

II. GPFN [5][6]

GPFN(Gaussian Potential Function Network)은 은닉층에 GPF(Gaussian Potential Function)라 부르는 유니트로 되어져 있다. x를 입력패턴으로 할 때 은닉층의 유니트 Ψ_i는 다음과 같이 정의한다.

$$\Psi_i = \Psi(x, p_i) = e^{-d(x,p_i)/2} \quad (1)$$

$$d(x, p_i) = d(x, m^i, K^i) = (x - m^i)^t K^i (x - m^i) \quad (2)$$

여기서, m^i와 K^i는 각각 i번째 GPF의 평균벡터와 공분산 행렬의 역행렬이다. $d(x, m^i, K^i)$를 다시 전개하면 식(3)와 같다.

$$d(x^i, m^i, K^i) = \sum_j \sum_k k^i_{jk}(x_j - m^i_j)(x_k - m^i_k) \quad (3)$$

여기서 x_j는 입력벡터 x의 j번째 요소, m^i_j는 평균벡터 m^i의 j번째 요소, k^i_{jk}는 역공분산 행렬 K^i의 (j, k)번째 요소를 각각 나타낸다. k^i_{jk}는 주변(Marginal) 표준편차 σ^i_j와 σ^i_k의 상관계수 h^i_{jk}로 나타낼 수 있다.

$$k^i_{jk} = \frac{h^i_{jk}}{\sigma^i_j \sigma^i_k} \quad (4)$$

여기서, σ^i_j는 양의 실수이다.

$$\begin{cases} h^i_{jk} = 1 & , j = k \\ h^i_{jk} \le 1 & , \text{그 외} \end{cases}$$

그림 1에 3층 GPFN의 구조를 나타내었다. 입력층과 출력층은 선형 유니트들로 구성되었으며, 은닉층은 GPF를 발생하는 GPFU (Gaussian Potential Function Unit)로 구성되었다.

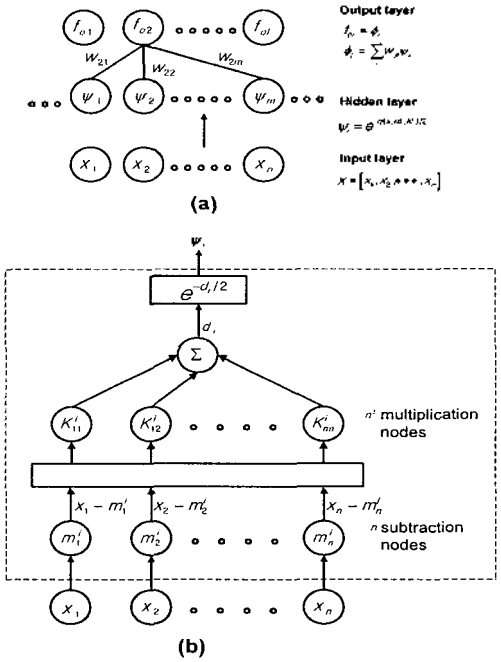

Fig. 1 GPFN의 개략도.
(a) GPFN,
(b) 입력층과 GPFU 사이의 연결

GPFN의 가중치 출력들은 원하는 Potential field를 합성하기 위해서 은닉층과 출력층 사이에 연결되어 더해진다. j번째 출력 유니트의 출력치를 ϕ_j로 하면

$$\phi_j = \sum_i w_{ji}\phi_i \tag{5}$$

로 된다.

이 네트워크를 학습하기 위해서는 역전파 학습법과 동일하게 오차함수에 최급하강법을 적용하고 교사패턴을 참조하여 파라미터를 갱신하는 것이다. p번째 교사패턴의 오차함수 E_p를 식(6)으로 정의한다.

$$E_p \equiv \frac{1}{2}\sum_{j=1}^{M}(t_{pj} - \phi_{pj}(\mathrm{n_j}))^2 \tag{6}$$

여기서 M은 출력 유니트의 수, t_{pj}는 목표 값을 나타낸다. ϕ_{pj}는 실제 출력의 j번째 요소를 나타내고, $\mathrm{n_j}$는 j번째 출력 유니트의 전 파라미터에서 열벡터로 된다.

$$\mathrm{n_j} \equiv [\mathrm{w_j^t}, \mathrm{m_j^t}, \sigma_j^t, \mathrm{h_j^t}]^t \tag{7}$$

방향벡터 $\Delta \mathrm{n_j} \equiv [\Delta \mathrm{w_j^t}, \Delta \mathrm{m_j^t}, \Delta\sigma_j^t, \Delta \mathrm{h_j^t}]^t$는 최급 하강법으로 구해진다. 그리고 파라미터의 갱신규칙은

$$\mathrm{n_j^{new}} = \mathrm{n_j^{old}} + \eta \Delta \mathrm{n_j} \tag{8}$$

로 된다. η는 학습률이다.

- j번째 출력층과 i번째 GPFU 사이의 연결강도

$$\Delta w_{ji} = -(t_j - \phi_j)\Psi_i \tag{9}$$

- 평균벡터 m^i의 j번째 요소

$$\Delta m_j^i = \sum_l k_{jl}^i (x_l - m_l^i)\Psi_i \sum_k (t_k - \phi_k) w_{ki} \tag{10}$$

- 주변(marginal) 표준편차 σ_j^i

$$\Delta\sigma_j^i = \sum_l k_{jl}^i \frac{(x_j - m_j^i)(x_l - m_l^i)}{\sigma_j^i} \Psi_i \sum_k (t_k - \phi_k) w_{ki} \tag{11}$$

- k_{jk}^i에 대한 공분산 계수 h_{jk}^i

$$\Delta h_{jk}^i = -\frac{1}{2}\frac{(x_j - m_j^i)(x_k - m_k^i)}{\sigma_j^i \sigma_k^i} \cdot \Psi_i \sum_k (t_k - \phi_k) w_{ki} \tag{12}$$

Ⅲ. RBFN

1. RBFN-Ⅰ

RBFN-Ⅰ은 음성의 각 구간에 대해서 HMM의 각 상태에 대한 확률분포 파라미터인 평균과 공분산을 RBFN의 중간층에서 가우시안 활성화 함수 출력에 이용한다. 즉, HMM을 5상태 4출력분포, 혼합수 2로 하였을 때 음성 구간을 출력분포 수 만큼의 구간으로 나누고 각 구간마다 혼합수 만큼의 중간층 유니트를 가지게 하는 것이다. 따라서 RBFN은 각 구간에서 중간층과 출력층 사이의 연결강도를 오차가 감소하도록 최급 강하법으로 학습한다. 이를 그림 2에 나타내었다. 그러므로 RBFN-Ⅰ은 GPFN처럼 가우시안 활성화 함수를 출력하기 위한 확률분포 파라미터를 추정하는 과정이 없기 때문에 학습시간을 대폭 단축할 수 있다.

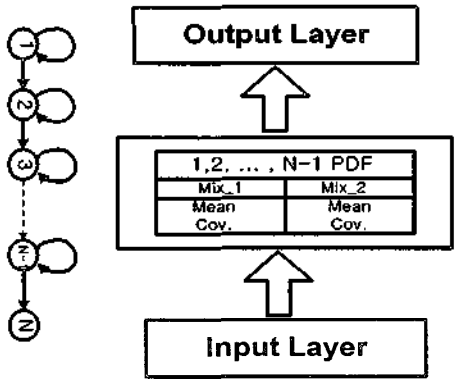

Fig. 2 RBFN-Ⅰ

2. RBFN-II

RBFN-Ⅱ는 음성의 각 구간에 대해서 HMM의 전체 상태에 대한 확률분포 파라미터, 즉 평균과 공분산을 RBFN의 중간층 가우시안 활성화 함수 출력에 이용한다. 즉, HMM을 5상태 4출력분포, 혼합수 2로 하였을 때 음성 구간을 출력분포 수만큼의 구간으로 나누고 각 구간마다 출력분포수×혼합수 만큼의 중간층 유니트를 가지게 하는 것이다. 따라서 RBFN은 각 구간에서 중간층과 출력층 사이의 연결강도를 오차가 감소하도록 최급 하강법으로 학습한다. 이를 그림 3에 나타내었다.

그러므로 RBFN-Ⅱ는 GPFN보다는 학습시간을 대폭 단축할 수 있지만 RBFN-Ⅰ보다는 학습시간은 더 소요된다.

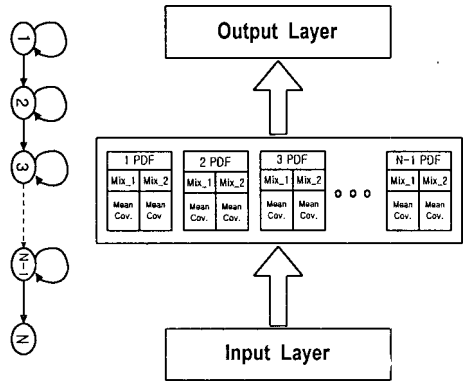

Fig. 3 RBFN-Ⅱ

Ⅳ. 실험 및 인식결과

실험에 이용된 음성 데이터는 단독 숫자음으로 ETRI의 샘돌이 데이터 중 "영, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구" 10개이며, 남성화자 20명이 4회 발성한 숫자음 중에서 처음 3회분은 학습용(600개)으로, 나머지 1회분은 평가용(200개) 데이터로 사용하였다. 그리고 표 1은 특징 파라미터를 추출하기 위한 분석 조건들이다.

RBFN의 구조는 단독 숫자음에서 예측차수를 1차로 하여 입력층과 출력층은 유니트

표 1. 특징파라미터의 추출

Filtering	LPF, 7kHz
A/D convert	16kHz, 16bit
Shifting Period	3.75ms
Window Length	16ms
Feature Parameter	10th order LPC Melcepstrum

수를 10개로 구성하여 한 프레임을 예측하도록 하였다. HMM은 혼합수를 2로 하고 상태수를 2~7로 증가시키면서 각 상태에 대한 확률분포를 구하고 이를 RBFN의 중간층의 가우시안 활성화 함수 출력에 이용한다. 예측형 GPFN은 HMM을 이용하지 않으므로 중간층의 유니트 수를 5, 10, 15, 20개로 하였다.

[방법 1] Predictive GPFN
[방법 2] Predictive RBFN-Ⅰ
[방법 3] Predictive RBFN-Ⅱ

방법 1은 HMM에 의해서 확률분포를 추정하지 않고 입력 데이터로부터 직접 확률분포를 구하고 이를 네트워크내에서 학습에 의해 재추정한다. 그러면 학습에 의해서 각 구간에 대한 확률분포가 추정되고, 중간층과 출력층 사이의 연결강도를 구한다.

그리고 방법 2와 3은 HMM에 의해서 추정된 확률분포 파라미터를 이용하기 때문에 네트워크에서는 이를 다시 추정하지 않고 중간층과 출력층 사이의 연결강도만 각 구간마다 학습한다. 따라서 방법 2는 HMM의 각 상태에 대한 확률분포를 이용하고, 방법 3은 전체 상태에 대한 확률분포를 이용하게 된다.

표 2는 방법 1, 2, 3으로 단독 숫자음에 대하여 실험한 인식결과이다. 그리고 그림 4와 5는 학습률과 상태수에 따른 RBFN-Ⅰ과 RBFN-Ⅱ

표 2. 단독 숫자음의 인식결과. [%]

방 법	학습	인식
HMM	100.0	99.0
MLP	98.0	96.0
J-RNN	98.5	97.5
E-RNN	98.7	98.0
GPFN	90.2	86.0
RBFN-Ⅰ	94.0	92.5
RBFN-Ⅱ	94.2	94.5

의 인식결과이다. 평가용 데이터를 기준으로 예측형 GPFN은 상태수 7, 중간층 유니트의 수가 20개일 때 86.0%로 가장 좋은 결과를 나타내었다. 그리고 RBFN-Ⅰ은 상태수가 4이고 학습률이 0.5일 때 92.5%로, RBFN-Ⅱ는 상태수가 4이고 학습률이 0.09일 때 94.5%로 가장 좋은 인식결과를 나타내었다. 이는 HMM의 99.0%과 예측형 RNN의 Elman망에서의 인식률 98.0%, Jordan망에서의 인식률 99.0%보다 4.5~6.5%정도 낮은 인식성능을 나타내었다.[7] 그러나 예측형 RBFN이 다른 네트워크에 비해서 학습횟수와 학습시간을 단축할 수 있었다.

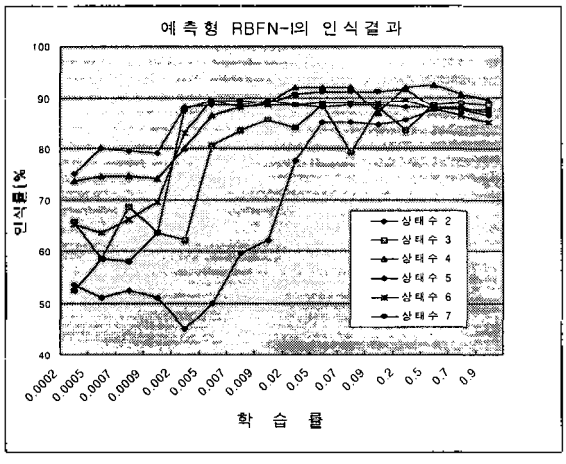

그림 4. 학습률에 따른 RBFN-Ⅰ의 인식결과

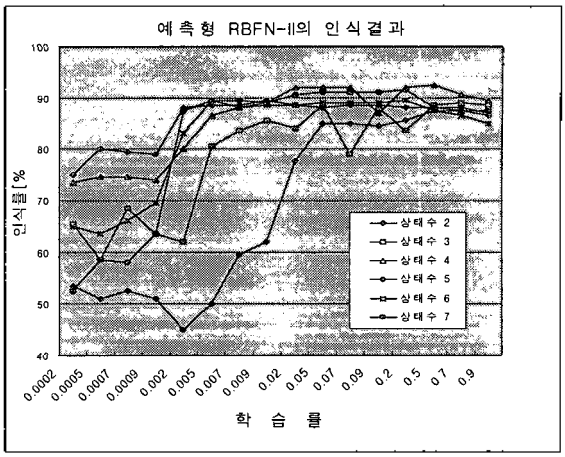

그림 5. 학습률에 따른 RBFN-Ⅱ의 인식결과

Ⅴ. 결 론

본 논문에서 제안한 예측형 RBFN은 HMM으로 추정한 확률분포 파라미터를 사용하여 중간층의 활성화 함수의 출력을 결정하고, 중간층과 출력층의 연결강도만 네트워크 내에서 학습한다. 그리고 HMM으로 추정한 확률분포 파라미터는 두 가지 방법으로 예측형 RBFN에 이용하였다. 첫 번째는 HMM의 각 상태의 혼합수만큼의 중간층 유니트를 주는 방법이고, 두 번째는 HMM의 혼합수×출력분포수 만큼의 중간층 유니트를 주는 방법이다.

실험결과에서 예측형 RBFN은 다른 방법들의 결과보다 4.5~6.5% 저하된 결과를 보였다. 그러나 다른 신경망에 비해서 전체 학습시간을 단축할 수 있었고 학습 반복 횟수를 10회로 작게 할 수 있는 장점이 있다.

따라서 RBFN의 인식성능을 향상시키기 위해서 음성패턴의 시간적 상관관계를 보다 잘 표현할 수 있는 RBFN의 구조에 대한 연구와 학습 알고리즘에 대한 연구가 필요하다고 판단된다.

본 논문에서 제안한 RBFN은 HMM과 신경망의 결합이라고 하는 향후 가장 유효한 방법의 하나로 사료되며 화자 적응화 시스템으로 확장할 수 있을 것으로 기대된다.

참고문헌

[1] K-F. Lee and H-W. Hon, "Large-vocabulary speaker-Independent Continuous Speech Recognition using HMM : The SPHINX System", Proc. ICASSP-88, pp.3-126, 1988.

[2] L. R. Bahl, et al. "Acoustic Markov Models used in the TANGORA speech recognition system", Proc. ICASSP-88, pp. 467-500, 1988.

[3] W. Huang, R. Lippmann, T. Nguyen, "Neural Nets for Speech Recognition", Conf. of the Acous. Society of America, Seattle WA, 1988.

[4] T. Kohonen et al., "Shift-Tolerant LVQ and Hybrid LVQ-HMM for Phoneme Recognition", in Speech Recognition, pp. 425-438, Morgan Kaufmann Publishers Inc. 1990.

[5] B. Kosko, *Neural Networks for Signal Processing*, pp.199-223, Prentice-Hall Internation, Inc.

[6] 김주성, 김수훈, 허강인, "RBFN을 이용한 음소인식에 관한 연구", 한국통신학회지, Vol. 22, No. 5, pp.1026-1035, 1997.

[7] 한학용, 김주성, 허강인, "회귀신경망을 이용한 음성인식에 관한 연구", 한국음향학회지, Vol. 18, No. 3, pp.40-48, 1999.