1999년도 한국음향학회 학술발표대회 논문집 제18권 제1(s)호

반연속 HMM과 RBF 혼합 시스템을 이용한 화자독립 음성인식에 관한 연구

문 연 주*, 전 선 도, 강 철 호
광운 대학교 전자통신공학과

A Study on Speaker-Independent Speech Recognition Using a Hybrid System of Semi-Continuous HMM and RBF

(Yun Joo Moon*, Sun Do June, Chul Ho Kang)
E-mail : pig@dspalpha.kwangwoon.ac.kr

요 약

본 논문에서는 기존의 반연속 HMM과 신경망 알고리즘인 RBF(Radial Basis Function)를 혼합한 형태를 음성인식에 적용한다.

기존의 반연속 HMM은 학습 과정에서 모든 모델과 상태에서 공유되는 L개의 가우시안 확률 밀도들과 각 가우시안 확률 밀도들의 가중치를 결정하는 혼합 밀도 계수에 의해 입력 음성의 특징을 확률적으로 모델링하는 혼합 확률을 얻고 또 Maximum likelihood와 Baum-Welch 알고리즘을 이용해 초기확률, 전이확률, 관측확률, 평균벡터 μ, 공분산 행렬 Σ을 학습해 나간다.

그러나 제안한 RBF/반연속 HMM 혼합형태는 RBF의 변형된 방식을 첨가해 반연속 HMM의 관측 파라미터를 RBF에 의해 결정함으로써 보단 분별력 있는 화자독립 인식 시스템이 된다. 그래서 인식 실험결과 인식률에 있어서 기존의 반연속 HMM보다 향상된 인식률을 얻는다.

Ⅰ. 서 론

현재 연구되고 있는 음성 인식 방법으로는 벡터 양자화, 신경망에 의한 인식, 시간적인 정합을 이용한 DTW(Dynamic Time Warping)알고리즘, 확률적인 방법으로 알려진 Hidden Markov Model(HMM)[4] 등이 있다. 그 중에서도 HMM은 음성의 시간적인 변이성을 통계적인 모델로 분석하므로 높은 인식률을 보이므로 신경망과 함께 결합하여 다양하게 연구가 되고 있다[8]. 인식방법에 있어서 화자종속의 경우 인식에 있어서는 별 문제가 안되지만 학습에 참여하지 단어을 인식시킬 경우에는 인식률이 저하되는 경향이 있다. 그래서 HMM과 신경망의 혼합시스템 방식으로 하여 각각 단어의 특징을 좀더 세분화시켜 학습을 시키므로 인식률 저하를 줄일 수 있다.

Ⅱ. 기존의 반연속 HMM 및 RBF

2.1 기존의 반연속 HMM

이산 HMM보다 작은 코드북을 사용하고 연속 HMM보다 적은 계산량이 필요하도록 두 경우를 결합한 것을 반연속 HMM (Semi-Continuous HMM;SCHMM)이라 한다[2][7].

이 경우는 벡터 양자화 코드워드를 가우시안 분포의 평균치들로 생각하며, 각 분포의 공분산 행열의 대각선값들을 코드북에 포함시키게 된다. 즉, 크기 L인 코드북에서 각 코드워드에 해당하는 D차 평균값 μ와 공분산 행열의 주대각선 성분 $\sum$가 D개 주어지게 된다[1]. 각 코드워드마다 공분산 행렬값이 주어지므로 일반적인 벡터 양자화의 경우와는 달리 유클리디안(Euclidean) 거리 대신 마할라노비스(Mahalanobis)거리를 사용하게 된다.

이와같은 경우, 확률 밀도 행열 B는 상태 j에서 1번째 코드워드에 해당하는 가우시안 성분을 발견할 상대적인

크기가 되므로 확률 밀도 행열 요소 b_{jl}은 연속 밀도 HMM의 C_{jf}(상태 j에서 f번째 가우시안 성분의 상대적인 크기) 와 같은 역할을 한다. 그러면 상태 j에서 관찰값 O_t를 발견할 확률은

$$p_j(o_t) = \sum_{l=1}^{L} b_{jl} p(o_t / \mu_1, \sum_l) \qquad (2\text{-}1)$$

로 주어진다.

이와같은 반연속 HMM을 이용하여 음성의 학습 데이터를 잘 표현하기 위해서는, 반연속HMM의 파라미터 재추정(parameter reestimation) 과정이 필요하다. 이것은 파라미터가 주어졌을 때, 관찰열을 발견할 확률을 반복적으로 최대화시키는 것으로서 EM(Expectation Maximization) 알고리즘이라 한다. 또한 주어진 반연속 HMM 파라미터들로부터 하나의 관찰열에 대응되는 가장 적합한 상태열을 찾는 방법으로 Viterbi 알고리즘이 있다.

반연속 HMM에서 재추정해야할 변수들은 행열, 상태천이 A행열, 출력확률행열 B 및 코드북의 μ, $\sum$ 값들이다. 이들은 다음의 재추정식으로부터 구한다.

$$\pi_i = \gamma_1(i) \qquad (2\text{-}2)$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \gamma_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \qquad (2\text{-}3)$$

$$b_j(k) = \frac{\sum_{t=1}^{T-1} \zeta_t(j,k)}{\sum_{t=1}^{T-1} \gamma_t(j)} \qquad (2\text{-}4)$$

$$\mu_j = \frac{\sum_{t=1}^{T-1} \zeta_t(j)\, o_t}{\sum_{t=1}^{T-1} \zeta_t(j)} \qquad (2\text{-}5)$$

$$\sum_j = \frac{\sum_{t=1}^{T-1} \zeta_t(i)(o_t - \mu_j)(o_t - \mu_j)^t}{\sum_{t=1}^{T-1} \zeta_t(j)} \qquad (2\text{-}6)$$

여기에서 중간 변수는 식 2-7, .2-8, 2-9, 2-10 이다.

$$\gamma_t(i,j) = P(s_t = i,\ s_{t+1} = j \mid O, \lambda) \qquad (2\text{-}7)$$

$$\gamma_t(i) = P(s_t = i, \mid O, \lambda) \qquad (2\text{-}8)$$

$$\zeta_t(i,k) = P(s_t = i,\ o_t - v_k \mid O, \lambda) \qquad (2\text{-}9)$$

$$\zeta_t(k) = P(o_t - v_k \mid O, \lambda) \qquad (2\text{-}10)$$

2.2 RBF 신경망 개요

RBF(Radial Basis Function) 신경망은 Broomhead와 Lowe에 의해 제안되었으며 그림 1와 같이 은닉층과 출력층의 2층으로 구성된 전방향의 지도학습 알고리즘을 갖는 신경망이다[2].

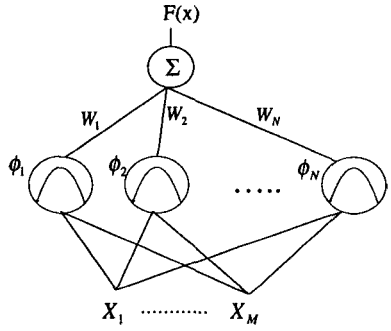

그림 1.RBF 신경망 구조

Figure 1.RBF network structure

기본적인 구조는 이층 퍼셉트론과 비슷하지만 이층 퍼셉트론과는 달리 은닉층은 단 한층으로만 구성되며 입력층과 은닉층 사이의 연결은 가중치를 갖지 않고 입력값을 그대로 받아들인다. 또 은닉층의 활성화 함수는 시그모이드 함수대신 여러 형태의 radial basis함수로 구성된다. 은닉층의 각 노드는 중심점(center)이라고 불리는 벡터를 포함하고 있는데 이 중심점들은 입력벡터 세트를 대표하는 벡터들이다.

일반적으로 RBF 신경망의 은닉층 출력은 주로 식 (2-11)와 같이 가우시안 함수를 활성화 함수로 하여 얻어진다.

$$\Phi_i(x) = \Phi(\|x - c_i\|^2 / \rho_i) = \exp(-\|x - c_i\|^2 / \rho_i) \qquad (2\text{-}11)$$

for i=1, 2, ... N

N은 중심점의 개수, c_i 는 RBF의 중심점, ρ_i 는 distance scaling 파라미터이다.

그리고 입력벡터와 출력사이의 비선형 함수의 근사는

식 (2-12)와 같이 비선형 basis 함수 Φ_i 의 선형조합으로 표현된다.

$$F(x) = \sum_{i=1}^{N} \omega_i \Phi_i(x) \qquad (2\text{-}12)$$

은닉층과 출력층 사이의 선형 가중치 ω_i 는 식 (2-13)과 같은 최소 자승 오차(LMS) 알고리즘을 이용하여 적응된다[5].

$$\Phi_i(x,t) = \exp(-\|x(t) - c_i(t)\|^2 / \rho_i),$$

$$e(t) = d(t) - \sum_{i=1}^{N} \omega_i(t-1)\ \Phi_i(x,t), \qquad (2\text{-}13)$$

$$\omega_i(t) = \omega_i(t-1) + g_\omega e(t)\ \Phi_i(x,t)$$

$$\text{for} \quad 1 \le i \le N$$

여기서 d(t)는 현재의 입력센터에 대한 목표출력, g_ω는 가중치에 대한 학습계수이다.

기본적인 RBF 신경망에서 중심점 c_i와 distance scaling(width) 파라미터 ρ_i 는 고정되며 단지 가중치 ω_i 만 적응시킨다. 실험적으로 충분한 수의 은닉층 노드를 갖고 중심점이 입력차원에 적절히 분포되어 있다면 RBF네트워크는 넓은 범위의 비선형 함수를 근사 시킬수 있다.

Ⅲ. 제안한 반연속과 RBF 혼합 시스템

3.1 학습 과정

이 논문에서 제안한 전체 시스템은 학습과정과 인식과정으로 나뉘어지는데, 먼저 학습과정 부분에서는 기존의 HMM과 마찬가지로 L개의 가우시안 확률 밀도들과 각 가우시안 확률 밀도들의 가중치를 결정하는 혼합 밀도 계수에 의해 입력 음성의 특징을 확률적으로 모델링하는 혼합확률을 얻어 Maximum likelihood와 Baum-Welch 알고리즘을 이용해 초기확률, 전이확률, 관측확률, 평균벡터 μ, 공분산 행렬 Σ을 학습해 나간다. 여기서 혼합 확률 밀도 계수(b)을 RBF의 Desired 값으로 해서 LMS 알고리즘의 적응화 과정을 통해 새로운 파라미터, 즉 가중치(Weight)를 구한다.

그림 2에 나타난 수식을 전개하면 다음과 같다.

$$G_1 = \sum_{t=1}^{T} X_t \ , \ G_2 = \sum_{t=1}^{T} X_t \ . \ . \ G_L = \sum_{t=1}^{T} X_t$$

X_t : 입력 데이터, G_l : 가우시안 함수 (3-1)

$$y_{s1} = \sum_{l=1}^{L} G_1 * W_{s1l} \ , \ y_{s2} = \sum_{l=1}^{L} G_l * W_{2l} \ \ . .$$

$$y_{sL} = \sum_{l=1}^{L} G_1 * W_{sLl} \qquad (3\text{-}2)$$

y_{sl} : 출력값 , W_{Sll} : 가중치

$$\varepsilon_{s1} = y_{s1} - d_{s1} \ , \varepsilon_{s2} = y_{s2} - d_{s2} \ . \ . \ . \ \varepsilon_{sL} = y_{sL} - d_{sL}$$

ε_{sl} : 에러값 (3-3)

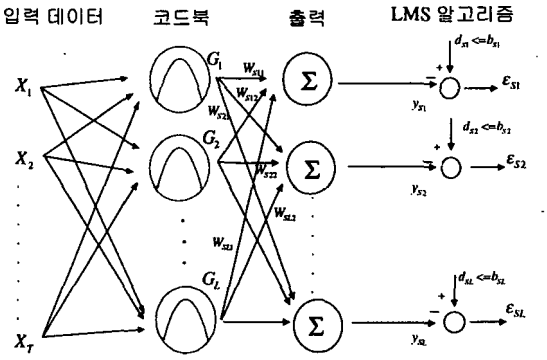

그림 2.제안한 RBF/HMM 혼합형 학습모식도

Figure 2. Learning block diagram of hybrid RBF/HMM

3.2 인식 과정

인식과정에서는 먼저 입력 데이터가 RBF과정를 거쳐 그 결과 나온 출력 값(y)이 Viterbi 인식 과정의 혼합 확률밀도 계수(b)로 사용된다.

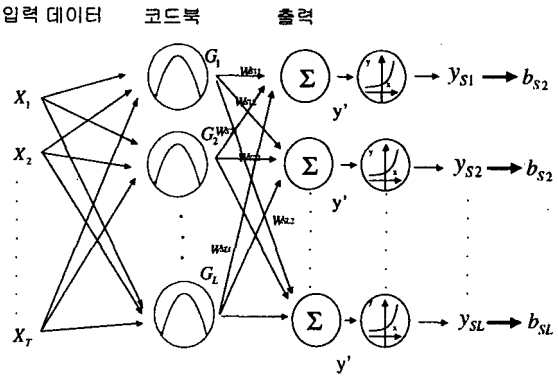

그림 3.제안한 RBF /HMM 혼합형 인식 모식도

Figure 3. Recognition block diagram of hybrid RBF/HMM

위 그림에 나타난 수식을 전개하면 다음과 같다.

$$G_1 = \sum_{t=1}^{T} X_t \quad , G_2 = \sum_{t=1}^{T} X_t \quad . \quad . G_L = \sum_{t=1}^{T} X_t$$

(3-4)

$$y'_{s1} = \sum_{l=1}^{L} G_1 * W_{s1l} \quad , \quad y'_{s2} = \sum_{l=1}^{L} G_l * W_{s2l} \quad \cdots$$

$$y'_{sL} = \sum_{l=1}^{L} G_1 * W_{sLl} \qquad (3\text{-}5)$$

y'_{sL} 은 양수와 음수 값으로 나타나는데 y'_{sL} 중에서 최소 값이 음수 값이므로 그 음수 값을 각각 성분에 대해 최소값을 빼서 y'_{sL} 값을 양수 값으로 정규화 시킨다.

$$y_{s1} = (\alpha * y'_{s1})^k, \quad y_{s2} = (\alpha * y'_{s2})^k \cdots$$

$$y_{sL} = (\alpha * y'_{sL})^k \quad , \quad \sum_{l=1}^{L} y_{sl} = 1 \qquad (3\text{-}6)$$

y_{sl} 의 모든 값을 더해서 1 이 되어야 하므로 y_{sl} 값을 모두 합해서 각각 성분에 나눈다. 그러면 확률 값을 얻을 수 있다.

Ⅳ. 실험 결과 및 고찰

4.1 실험데이타 및 인식 모델

본 실험에 이용된 음성 데이터의 Preemphasis는 $H(z) = 1 - 0.95z^{-1}$ 이고 Frame Blocking 은 256 speech sample을 128개씩 shift하고 14차 LPC 캡스트럼 계수를 이용하여 k-means 알고리즘을 이용해 64개의 코드워드로 구성된 코드북 벡터를 생성하였다.

학습에 사용된 단어는 10명의 남성화자가 한국어 10개 격리 단어 지역명을 두 번씩 반복 발음하였다.

또한 실험에 사용된 인식 모델은 반연속 HMM이고 반연속 HMM의 상태 갯수를 10개로 하고 left to right 모델이 적용되었다.

4.2 실험결과 및 고찰

본 논문에서 학습에 참여한 10명과 학습에 참여하지 않은 5명을 인식 실험에 테스트하였다.

표 1은 기존의 반연속 HMM의 인식률 결과를 나타내고 표 2는 제안된 BRF/HMM의 인식률을 나타낸다.

표에서 나타나듯이 제안된 방식이 기존의 방식보다 인식률이 향상 되었다.

표 1. 기존의 반연속 HMM 인식률

	인식률(%)	
학습화자	첫 번째발음	93
	두 번째발음	94
비학습화자	첫 번째발음	78
	두 번째발음	80

표 2. 제안된 방법의 인식률

	인식률(%)	
학습화자	첫 번째발음	95
	두 번째발음	94
비학습화자	첫 번째발음	86
	두 번째발음	84

Ⅴ. 결 론

본 연구에서는 기존의 HMM과 신경망과 HMM의 하이브리드 형태에 대해 음성 인식 실험을 하였다.

화자 종속인 경우보다 화자독립인 경우에 인식률이 더 향상됨을 볼 수 있었다.

현재의 인식 실험 시 단어의 학습이 최적화 되지 않은 상태이어서 화자독립 테스트에서 인식률이 화자 종속보다 상당히 인식률이 감소 됨을 알 수 있다. 따라서 학습 정도와 더불어 각각 단계에서 임계치 값을 어떻게 설정하는냐가 앞으로 추후과제로 남아있다.

Ⅵ. 참고문헌

[1] Alberto Leon-Garcia. "Probability and Random Processes for Electrical Engineering."

[2] Simon Haykin. "Neural Networks."

[3] X.D Huang, Y. Ariki, M.A. Jack. "Hidden Markov Models for Speech Recognition."

[4] Lawrence Rabiner, Biing-Hwang Juang. "Fundamentals of Speech Recognition."

[5] Bernard Widrow, Samuel D. Stearns. "Adaptive Signal Processing."

[6] X.D. Huang, M.A. Jack. "Semi-continuous hidden Markov models for speech recognition." Computer Speech and Language, vol.3, pp. 239-251, 1989.

[7] X.D. Huang, "Semi-continuous hidden Markov models for speech recognition," Ph.D. thesis, Department of Electrical Engineering, University of Edinburgh, 1989.

[8] Eliot Singer, Richard P. Lippmann, "A Speech Recognition Using Radial Basis Function Neural Networks In An HMM Framework," Proc. ICASSP, pp.629-632, 1992.