회귀신경망 예측 HMM을 이용한 음성 인식에 관한 연구

박경훈, 한학용, 김수훈, 허강인
동아대학교 전자공학과

A study on Speech Recognition Using Recurrent Neural Predictive HMM

Park Kyung Hoon, Han Hak Yong, Soo Hoon Kim, Kang In Hur
Dept. of Electronics, Dong-A University
E-mail : kihur@mail.donga.ac.kr

요약

본문에서는 예측형 회귀신경망과 HMM의 하이브리드 네트워크인 회귀신경망 예측 HMM을 구성하였다. 회귀신경망 예측 HMM은 예측형 회귀신경망을 HMM의 각 상태마다 예측기로 정의하여 일정치인 평균벡터 대신에 과거의 특징벡터의 영향을 받아 동적으로 변화하는 신경망에 의한 예측치를 이용하므로 학습패턴 설정자체가 시변성을 반영하는 동적 네트워크의 특성을 가진다. 따라서 음성과 같은 시계열 패턴의 인식에 유리하다. 회귀신경망 예측 HMM은 예측형 회귀신경망의 구조에 따라 Elman망 예측 HMM과 Jordan망 예측 HMM으로 구분하였다. 실험에서는 회귀신경망 예측 HMM의 상태수를 4, 5, 6으로 증가시켜 각 상태 수별로 예측차수 및 중간층 유니트 수의 변화에 따른 인식성능을 조사하였다. 실험결과 평가용 데이터에 대하여 Elman망 예측 HMM은 상태수가 6이고, 예측차수가 3차, 중간층 유니트의 수가 15차원일 때, Jordan망 예측 HMM의 경우 상태수가 5이고, 예측차수가 3차, 중간층 유니트의 수가 10차원일 때 각각 98.5%로 우수한 결과를 얻었다.

I. 서론

현재 음성인식의 대표적인 방법중의 하나 HMM 법은 개인차나 조음결합 등으로 나타나는 음성패턴의 변동이 정확하게 반영되고 확률 통계론에 의한 이론적 전개가 용이하며 음소나 음절 단위의 모델을 단어, 문장 등의 단위로 쉽게 확장할 수 있는 장점이 있다. 신경망은 화자의 개인차 등에 의한 스펙트럼의 변동을 유니트간의 결합 가중치로서 표현할 수 있고 한번에 많은 프레임의 데이터를 입력할 수 있는 장점이 있다. 회귀 신경망은 출력층과 중간층의 활성화 값을 회귀시켜 신경망에 동적특성을 부여하고 있으며 Elman망과 Jordan망이 대표적이다.

HMM과 신경망의 장점을 함께 사용할 수 있는 하이브리드 네트워크에 대한 연구 또한 진행되고 있다. 본 논문에서는 이러한 점을 고려하여 HMM과 예측형 회귀신경망의 장점을 함께 사용할 수 있는 하이브리드 네트워크인 회귀신경망 예측 HMM을 구성하였다. 실험에서는 단독 숫자음에 대한 회귀신경망 예측 HMM의 인식실험 결과를 CHMM과 예측형 회귀신경망에 의한 인식결과와 비교 · 검토하였다.

II. 연속 출력분포 HMM

Left-to-right형 HMM은 그림 1과 같은 유한 오토마타로 정의된다.

HMM을 이용한 음성인식의 경우는 먼저 인식에 필요한 수 만큼의 표준패턴을 학습해 두고 입력패턴에 대하여 그 출력확률이 최대가 되는 표준패턴을 인식결과로 한다.

연속 출력분포 HMM은 상태 i에서 j로의 천이확률 a_{ij} 및 천이경로에서 심벌 k의 출력확

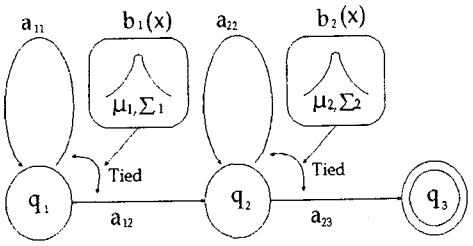

그림 1. 연속 출력확률 HMM

률 b_{ijk}를 학습 데이터에서 구하기 위한 Baum-Welch 알고리즘은 다음과 같다. 상태수를 N, T를 심벌계열의 길이, 전향확률을 $\alpha(i,t)$ $(i=1,2,\cdots,N;\ t=1,2,\cdots,T)$라 하고, 후향확률을 $\beta(j,t)$ $(j=1,2,\cdots,N;\ t=T,T-1,\cdots,0)$, 모델 M의 심벌 계열 $o = o_1\ o_2\ \cdots\ o_T$를 출력하는 확률을 $P(o \mid M)$, 상태 i에서 상태 j로의 천이가 시각 t에서 발생할 확률을

$$\gamma_t(i,j) = \frac{\alpha(i,t-1)a_{ij}b_{ij}(o_t,\mu_{ij},\Sigma_{ij})\beta(j,t)}{P(o \mid M)} \quad (1)$$

로 정의하면 천이확률의 추정식은

$$a_{ij} = \sum_t \gamma_t(i,j) / \sum_t\sum_j \gamma_t(i,j) \quad (2)$$

$$b_{ijk} = \sum_{t:\,o_t=k} \gamma_t(i,j) / \sum_t \gamma_t(i,j) \quad (3)$$

와 같고, 출력벡터 o_t 가 n차원의 정규분포에 따른다고 가정할 수 있는 경우 출력확률 밀도함수는

$$b_{ij}(o_t,\ \mu_{ij},\ \Sigma_{ij}) = \frac{\exp\{-(o_t-\mu_{ij})'\Sigma_{ij}^{-1}(o_t-\mu_{ij})/2\}}{(2\pi)^{n/2}|\Sigma_{ij}|^{1/2}} \quad (4)$$

로 주어진다. 여기서, μ_{ij}는 출력벡터의 평균치, Σ_{ij}는 공분산행렬, t는 전치, -1은 역행렬을 나타낸다. 여기서 μ_{ij}, Σ_{ij}의 추정식은 다음 식으로 주어진다.

$$\mu_{ij} = \sum_t \gamma_t(i,j)o_t / \sum_t \gamma_t(i,j) \quad (5)$$

$$\Sigma_{ij} = \frac{\sum_t \gamma_t(i,\ j)(o_t-\mu_{ij})(o_t-\mu_{ij})^t}{\sum_t \gamma_t(i,\ j)} \quad (6)$$

III. 예측형 회귀신경망

예측형 신경망은 과거의 특징벡터로부터 다음 시각의 벡터를 예측하므로 서로 인접한 특징벡터 사이의 상관 관계를 잘 반영할 수 있다.

회귀신경망 예측 HMM에서는 HMM의 각 상태마다 신경망에 의한 예측기를 정의하여 시각 t에서의 입력벡터 y_t와 각 상태 i에 대응하는 네트워크가 출력하는 예측벡터 y_{it}'와의 차를 HMM 으로 학습하는 것이다. 연속출력 확률분포 HMM의 출력밀도 함수 $b_{ij}(y_t,\ \mu_{ij},\Sigma_{ij})$ 대신에 $b_{ij}(y_t,\ y_{it}',\ \Sigma_{ij})$을 사용하므로 일정치인 평균벡터 대신에 과거의 특징벡터의 영향을 받아 동적으로 변화하는 신경망에 의한 예측치를 이용하는 것이 된다. 따라서 식 (1)과 식 (4) 대신 각각

$$b_{ij}(y_t,\ y_{it}'\ \Sigma_{ij}) = \frac{\exp\{-(y_t-y_{it}')'\Sigma_{ij}^{-1}(y_t-y_{it}')/2\}}{(2\pi)^{n/2}|\Sigma_{ij}|^{1/2}} \quad (7)$$

$$\gamma_t(i,j) = \frac{\alpha_{t-1}(i)a_{ij}b_{ij}(y_t,y_{it}',\Sigma_{ij})\beta(j)}{p(Y|M)} \quad (8)$$

을 이용하여 분산을 다음과 같이 재 추정한다.

$$\Sigma_{ij}' = \frac{\sum_t \gamma_t(i,j)(y_t-y'_{it})(y_t-y'_{it})^t}{\sum_t \gamma_t(i,j)} \quad (9)$$

y_{it}'는 신경망으로 추정하므로 Baum-Welch 알고리즘과 BP 알고리즘을 조합하여 학습을 행하는 것이 된다. 회귀신경망 예측 HMM은 3층 구조의 예측형 회귀신경망과 CHMM으로 구성되어진다. 그림 2는 Elman망 예측 HMM의 구조이며 중간의 유니트가 문맥층으로 회귀되는 예측형 Elman망의 출력층은 HMM의 각 상태에 대응된다. 그림 3은 Jordan망 예측 HMM의 구조이며 출력층의 유니트가 문맥층으로 회귀되는 예측형 Jordan망 출력층은 HMM의 각 상태에 대응된다.

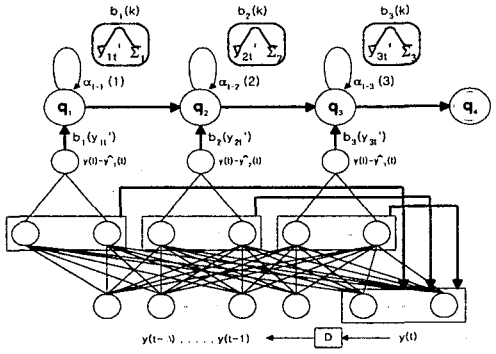

그림2. Elman망 예측HMM

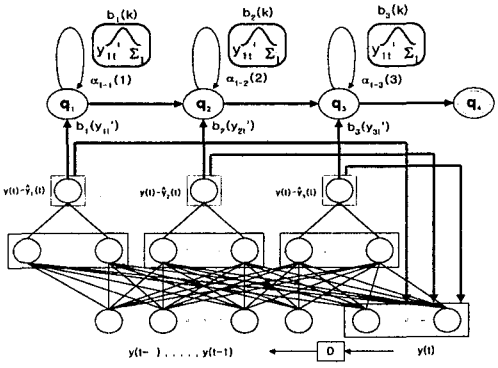

그림 3. Jordan망 예측HMM

IV. 인식 실험

4.1 음성데이터 및 분석조건

실험에서 사용한 음성데이터는 ETRI의 샘돌이 데이터중 단독 숫자음 "영"에서 "구"까지 10개이며, 남성화자 20명이 4회 발성한 총 800개의 데이터중 처음 3회분은 학습용(600개)으로, 나머지 1회분은 평가용(200개) 데이터로 사용하였으며 분석조건은 16KHz, 16Bit 샘플링하여 16ms의 분석창으로 프레임간격 3.75ms로 단구간 분석하여 10차 LPC Melcepstrum을 특징파라메터로 사용하였다.

4.2. 실험결과 및 고찰

실험에서는 단독 숫자음에 대하여 회귀신경망 예측 HMM의 상태수를 4, 5, 6으로 증가시켜 각 상태 수별로 예측차수 및 중간층 유니트 수의 변화에 따른 인식성능을 비교 · 검토하였다. 예측차수가 2차, 3차, 4차인 경우 회귀신경망 예측 HMM의 입력층 유니트의 수는 각각 20개, 30개, 40개이고 출력층은 10차원이다. 문맥층은 Elman망 예측 HMM의 경우 중간층 유니트의 수가 10차원일 때 10개, 15차원일때 15개, 20차원일 때 20개이다. Jordan망 예측 HMM의 경우 문맥층은 출력층 유니트의 수와 같은 10개이다. 회귀신경망 예측 HMM의 학습과정은 먼저 예측형 회귀신경망을 BP 알고리즘을 이용하여 학습하고 신경망에서 구한 파라메터 값을 이용하여 HMM을 Baum-Welch 알고리즘에 의하여 학습한다. HMM의 재 추정에 의한 최대학습 회수는 10회로 제한하였고 신경망은 600회 학습하였다.

표1은 CHMM의 각 상태수에 따른 인식결과이고, 표2는 예측형 회귀신경망의 예측차수 및 중간층 유니트 수의 변화에 따른 인식률이다. 표3와 표4은 각각 Elman망 및 Jordan망 예측HMM의 상태수, 예측차수 그리고 중간층 유니트수의 변화에 따른 인식률이다.

표1. CHMM의 인식률

	인식률(%)	
상태수	학습	평가
4	98.7	98.5
5	100.0	99.0
6	99.3	98.0

실험결과 전반적으로 회귀신경망 예측 HMM의 상태수가 4에서 5, 6으로 증가됨에 따라 인식률이 향상되었고, 예측차수 및 중간층 유니트 수의 변화보다는 상태수 변화에 크게 반응하였다.

표2. 예측형 회귀신경망의 인식률

예측차수	중간층 유니트수	인식률(%)			
		Elman망		Jordan망	
		학습	평가	학습	평가
2차	10	96.8	94.0	97.2	93.0
	15	98.7	95.5	98.7	97.0
	20	99.0	95.0	99.2	97.5
3차	10	97.2	95.5	98.3	95.5
	15	99.0	95.5	98.5	97.0
	20	99.0	94.0	99.2	97.0
4차	10	98.2	97.5	97.2	94.5
	15	98.3	95.0	99.5	95.0
	20	99.3	95.0	99.7	98.5

표3. Elman망 예측 HMM의 인식률

예측 차수	중간층 유니트 수	인식률(%)					
		4 상태		5 상태		6 상태	
		학습	평가	학습	평가	학습	평가
2차	10	96.3	95.0	98.0	97.0	99.0	98.0
	15	97.0	95.5	99.2	98.0	99.0	97.5
	20	96.7	96.5	98.8	97.0	98.2	96.5
3차	10	96.2	94.5	98.7	97.0	99.5	96.5
	15	97.0	96.5	98.3	97.0	99.0	98.5
	20	97.2	96.0	98.2	97.0	98.8	97.0
4차	10	95.8	95.0	99.0	98.0	99.5	97.0
	15	96.5	96.0	97.5	98.0	98.8	95.0
	20	96.2	96.0	98.2	96.5	99.2	98.0

표4. Jordan망 예측 HMM의 인식률

예측 차수	중간층 유니트 수	인식률(%)					
		4 상태		5 상태		6 상태	
		학습	평가	학습	평가	학습	평가
2차	10	97.3	96.5	98.7	97.0	98.8	98.0
	15	97.0	95.5	98.3	97.5	98.2	97.0
	20	96.8	96.0	97.8	96.5	98.5	97.5
3차	10	96.0	95.5	98.5	98.5	99.3	96.5
	15	96.8	97.0	98.2	96.5	98.8	98.0
	20	97.3	96.0	98.8	97.0	99.3	96.5
4차	10	97.3	95.5	98.3	98.0	99.0	97.0
	15	96.8	96.5	98.8	98.0	99.2	97.0
	20	95.8	95.0	99.2	97.5	99.3	98.0

그림 4는 회귀신경망 예측 HMM의 실험결과 중 상태수가 5일 때 예측차수 및 중간층 유니트 수의 변화에 따른 숫자음 "사"의 인식률이 가장 나쁘며, 주로 "삼"으로 오인식 하였다. 그리고 "일"의 경우 "칠"로, "이"의 경우 "일"로, "오"의 경우 "구"로, "구"의 경우 "오"로 오인식이 발생하였다.

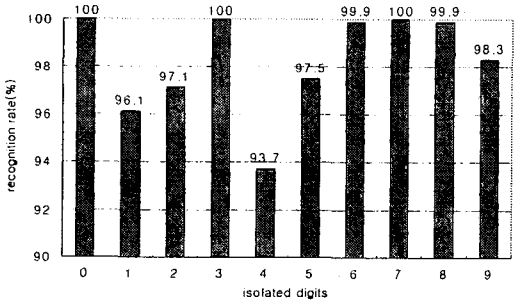

그림 4. 숫자음의 인식결과

V. 결론

본문에서는 예측형 회귀신경망과 HMM의 장점을 함께 사용할 수 있는 하이브리드 네트워크인 회귀신경망 예측 HMM을 구성하였다. 실험에서는 단독 숫자음에 대하여 회귀신경망 예측 HMM의 상태수를 4, 5, 6으로 증가시켜 각 상태수별로 예측차수 및 중간층 유니트 수의 변화에 따른 인식성능을 조사하였다. 또한 위의 결과를 CHMM 및 예측형 회귀신경망과 인식성능을 비교하였다.

실험결과 평가용 데이터에 대하여 Elman망 예측 HMM은 상태수가 6이고, 예측차수가 3차, 중간층 유니트의 수가 15차원일 때 98.5%, Jordan망 예측 HMM은 상태수가 5이고, 예측차수가 3차, 중간층 유니트 수가 10차원 일 때 98.5%의 인식률을 보였다. 회귀신경망 예측 HMM은 예측차수와 중간층의 증가에 따른 인식률의 변화보다는 상태수 변화에 크게 반응하는 결과를 보였다.

참고 문헌

1. Ken-ichi Iso and Takao Watanabe, "Speaker-Independent Word Recognition Using A Neural Prediction Model," proc. ICASSP'90, pp.441-444, 1990.

2. J.Tebelskis, "Speech Recognition using Neural Networks", a thesis for doctorate, Carnegie Mellon University, 1995.

3. Nile L.T. and Silverman H.F.: "Combining hidden Markov model and neural network classifiers", proc. int. conf. ASSP, pp.417-420, 1990.

4. K.Hassanein, L.Deng, M.I.Elmasry, "Vowel Classification Using A Neural Predictive HMM:ADiscriminative Training Approach", proc. ICASSP'94, pp. II-665-668, 1994.

5. 김수훈, 이종진, 허강인, "이산 연속분포 HMM을 이용한 연속음성 인식", 한국음향학회 논문지, 제14권 제1호, pp.81-89, 1995.

6. S.H.Kim, S.B.Kim, S.Y.Koh, K.I.Hur: "A Study on the Syllables Recognition Using Neural Network Predictive HMM", the Journal of the Acoustical Society of Korea, Vol. 17, No.2E, pp.26-30, 1998.

7. S.H.Kim, S.Y.Koh, K.I.Hur: "A Study on the Recognition of the isolated Digits Using Recurrent Neural Predictive HMM", TENCON'99 Vol. I, pp.593-596, 1999.