Hidden Markov Network를 이용한 음향학적 음소모델 작성에 관한 검토

°오 세 진, 임 영 춘, 황 철 준*, 김 범 국*, 정 현 열
영남대학교 전자정보공학부
*대구과학대학 정보전자통신계열

A Study on Construction of Acoustical Phoneme Models Using Hidden Markov Network

°Se-Jin Oh, Young-Choon Lim, Cheol-Jun Hwang*, Bum-Koog Kim*, Hyun-Yeol Chung
School of Electrical Eng., & Computer Science, Yeungnam University
*Informational Electronics & Communication Div., Taegu Science College

요 약

본 논문에서는 음성인식 시스템의 음향모델 개선을 위한 기초적 연구로서, 문맥적인 요소를 필요로 하는 SSS(Successive State Splitting)와 필요로 하지 않는 SSS-free 알고리즘을 이용한 HMnet(Hidden Markov Network) 음향모델 작성방법에 대해 검토하고 작성한 음향모델을 한국어에 적용하여 그 유효성을 확인하였다. HMnet을 이용한 음소모델의 작성방법은 전체 학습 데이터에 대해서 각각 2개의 상태를 가지는 초기 모델을 작성한 후, 이를 시간과 문맥방향으로의 최대 분포를 가지는 상태를 재분할한 후 임의의 상태수가 될 때까지 상태분할을 계속적으로 수행케 하여 각 음소모델을 작성하게 된다. 작성한 HMnet 음향모델의 유효성을 확인하기 위해 ETRI 445 단어의 3인에 대한 화자종속 음소인식 실험을 수행하였다. 인식실험 결과, SSS 알고리즘을 이용한 화자종속실험의 경우 상태수 520에서 평균 62.8%의 인식률을, SSS-free 알고리즘의 경우 상태수 420에서 평균 64.2%의 인식률을 얻었다. 이 결과는 HMM을 이용한 경우(약43.4%)보다 20%이상의 인식률 향상을 보여 이 알고리즘의 유효성을 확인할 수 있었다. SSS와 SSS-free를 비교한 경우, SSS-free가 SSS보다 낮은 상태수에서 평균 1.4% 향상된 인식률을 보였다.

1. 서 론

음성인식을 위한 인식단위로서 확장성, 훈련성, 대용량성 등을 고려하여 유사음소 단위가 많이 이용되고 있다[1]. 이 경우, 유사음소는 조음결합 등의 영향을 크게 받기 때문에 하나의 모델로서 표현하는데는 한계가 있다. 따라서 음소의 음향학적 특성을 변화시키는 요인(선행음소와 후행음소)까지 고려한 이음(allophone)을 인식 단위로 하는 방법이 소개되고 있으며 그 유효성이 확인되고 있다[2]. 그러나 이음을 인식단위로 할 경우에는 음소의 경우와 비교하여 모델의 수가 크게 증가하기 때문에 학습샘플의 수에 제한이 있는 경우에는 모델을 학습하는데 큰 문제가 될 수 있다.

따라서 신뢰성이 높은 모델을 구성하기 위해서는 음소환경을 분할하여 하나의 모델마다 학습 샘플의 수를 감소시키지 않고 모델을 학습하는 연구가 필요하다. 이러한 인식단위를 적절하게 설정하는 방법으로는 인간의 음성학적 지식에 근거한 방법과 주어진 샘플에 대해 음소환경 공간을 분할하는 방법, 전체 이음모델을 학습하고 음향학적으로 유사한 상태를 공유하는 방법 등이 제안되고 있다[5]. 그러나 이러한 방법들은 경험에 의한 인식시의 척도(우도)와는 다른 척도에 의해 음향모델의 구조가 결정되기 때문에 인식률이 최대가 되도록 구성하는 것은 어려운 일이다. 일반적인 HMM에서도 상태를 연결하고 모델의 구조를 결정하는데는 경험에 의한 경우가 많이 있다.

따라서 본 논문에서는 이러한 문제점을 해결하기 위해 제안된 SSS 및 SSS-free 알고리즘을 이용하여 모델의 구조를 자동으로 결정하고 모델의 음향학적 특징을 포함하는 HMnet 음소모델을 검토한 후 한국어에 적용하여 그 유효성을 검토한 결과를 보고한다.

2. Hidden Markov Network(HMnet)

그림 1에 일반적인 HMnet[3]의 구조를 나타내었다. 이 HMnet은 여러 개의 상태를 네트워크로 연결한 것으로, 각 상태는 다음과 같은 정보를 가지고 있다.

- 상태 번호
- 받아들여질 수 있는 음소환경 카테고리
- 선행 상태와 후속 상태 리스트
- 출력확률 분포의 파라미터
- 자기천이 확률과 후속 상태로의 천이 확률

HMnet은 HMM과 유사하나 고정된 상태수를 가지는 HMM과는 달리 음향학적 특징에 따라 각 음향모델이 다양한 상태수를 가진다는 점이 다르다. 그러나 각 음향모델이 단일 네트워크로 구성한 경우에는 HMM과 같이 취급할 수 있다.

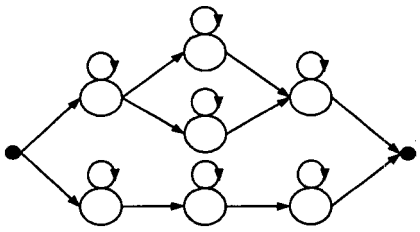

그림 1. HMnet의 구조

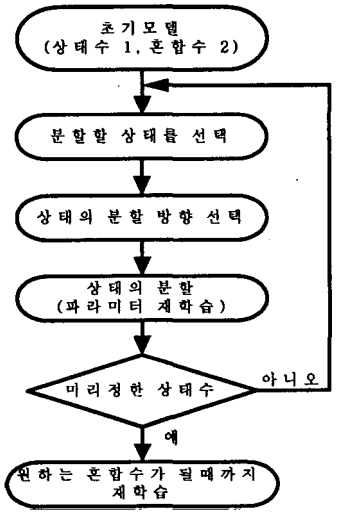

그림 2. SSS 및 SSS-free 알고리즘 구성도

3. 상태분할 알고리즘

최적의 HMnet을 구성하기 위해서는 음소환경 카테고리의 분류와 상태공유 구조 등에 관한 조합이 발생하는 문제를 해결할 필요가 있지만 이것을 실제로 구성하는 것은 어려운 일이다. 따라서 본 논문에서는 이를 해결하기 위해, [3]과 [4]에서 제안한 SSS 알고리즘과 SSS-free 알고리즘을 이용하였다. 본 논문에서 이용한 자동으로 모델의 상태를 분할하는 방법을 그림 2에 나타내었다. 이하 두 알고리즘을 이용한 모델의 학습에 관해서 간략히 설명한다.

1. 초기 모델의 학습

▪ SSS의 경우

초기모델로서 하나의 상태에 출력분포가 단일 가우스 분포(대각 공분산 행렬)를 가지는 HMM을 이용하여, 전체 학습 샘플에 대해 학습한다.

▪ SSS-free의 경우

초기모델로서 하나의 상태에 출력분포가 단일 가우스 분포(대각 공분산 행렬)를 가지는 HMM을 각 음소마다 이용하여, 전체 학습 샘플에 대해 학습한다.

2. 분할할 상태의 결정(SSS, SSS-free 공통)

모든 상태 중에서 출력분포가 가장 큰 상태를 분할할 상태로 선택한다. 출력분포는 단일 가우스 분산 값과 추정에 이용한 샘플 수를 곱한 것을 기준으로 한다. i번째 상태에서 출력분포 크기 d_i는 식 (1)을 이용하여 계산한다.

$$d_i = n_i \times \sum_k^K \frac{\sigma_{ik}^2}{\sigma_{Tk}^2} \qquad (1)$$

$$\sigma_{ik}^2 = \lambda_{i1}\sigma_{i1k}^2 + \lambda_{i2}\sigma_{i2k}^2 + \lambda_{i1}\lambda_{i2}(\mu_{i1k} - \mu_{i2k})^2$$

여기서, K는 파라미터의 차원, λ_{i1}, λ_{i2}는 상태 i의 가중계수, μ_{i1k}, μ_{i2k}는 상태 i의 k번째 평균, σ_{i1k}^2, σ_{i2k}^2는 상태 i의 k번째 분산, n_i는 상태 i에 대한 학습 샘플의 수, 그리고 σ_{Tk}^2는 모든 샘플의 k번째 분산을 나타낸다.

3. 상태 분할

선택된 상태를 다시 두 단계로 분할하게 되는데 새로운 상태의 출력확률 분포를 아래와 같이 구하게 된다.

(a) 분할할 상태를 통해 전체 학습 샘플에 대해 Viterbi 알고리즘을 사용하여 상태가 출력하는 샘플의 부분 계열을 추출하게 된다.

(b) (a)에서 추출한 모든 학습 샘플의 부분계열을 이용하여 1상태, 2혼합의 HMM을 학습한다.

(c) 구해진 2개의 가우스 분포를 각각의 새로운 상태에 할당한다.

이와 같이 새로운 상태의 출력확률 분포를 구한 후, 새로운 상태의 위치를 시간 방향(직렬)으로 연결

한 경우의 학습 샘플에 대한 우도 P_t와 문맥 방향(병렬)으로 연결한 경우의 우도 P_c를 계산하고 우도가 높은 것을 선택하게 된다. P_t와 P_c는 아래와 같은 과정을 통해 계산된다.

(1) 문맥방향으로의 상태분할

문맥방향으로의 분할은 2가지 경로를 고려할 수 있는데 이때 각각의 학습 샘플이 어느 쪽의 상태를 선택할 지를 결정할 필요가 있다.

• SSS의 경우

각각의 학습 샘플에 대해 문맥환경 요소(선행음소와 후행음소)로 그룹을 나누고 이 그룹에서 우도가 높은 상태를 결정하게 된다. 이때 상태 결정은 식 (2)를 이용한다.

$$P_c = \max_j \sum_l \max(P_m(y_{jl}), P_M(y_{jl})) \qquad (2)$$

여기서, j는 상태에서 문맥환경 요소를, y_{jl}은 요소 j의 값이 l번째 요소가 되는 학습 샘플의 부분집합을, $P_m(y_{jl})$는 y_{jl}을 상태 m에 할당할 때의 우도를, $P_M(y_{jl})$는 y_{jl}을 상태 M에 할당할 때의 우도를 각각 나타낸다.

문맥환경 요소 j를 결정한 후 식 (3)을 이용하여 분포 e_{jl}을 결정하게 된다.

$$\begin{cases} e_{jl} \in E_{m_j} & (P_m(y_{jl}) \geq P_M(y_{jl})) \\ e_{jl} \in E_{M_j} & (P_m(y_{jl}) < P_M(y_{jl})) \end{cases} \qquad (3)$$

여기서, e_{jl}은 문맥환경 요소 j에 속하는 l번째 성분을, E_{m_j}는 상태 m을 통과하는 성분들의 집합을, E_{M_j}는 상태 M을 통과하는 성분들의 집합을 각각 나타낸다.

• SSS-free의 경우

각각의 학습 샘플에 대해서 우도가 높은 상태를 선택하게 되며 이때 상태 결정은 식 (4)를 이용한다.

$$P_c = \sum_{y_j \in Y_m} \max(P_m(y_j), P_M(y_j)) \qquad (4)$$

여기서, Y_m은 분할할 상태 m을 통한 학습 샘플의 집합을, y_j는 상태 m을 통한 j번째 학습 샘플을, $P_m(y_j)$는 y_j를 상태 m에 할당할 때의 우도를, $P_M(y_j)$는 y_j를 상태 M에 할당할 때의 우도를 각각 나타낸다.

(2) 시간방향으로의 상태분할

시간방향으로의 상태 분할은 2개의 상태를 직렬로 연결하여 학습한다. 여기서는 위치의 순서에 따라 두 가지의 가능성이 존재하며 각각의 우도를 계산한 후 높은 우도를 가지는 상태의 우도 P_t를 선택한다.

4. 분포의 재추정(SSS, SSS-free 공통)

이 시점에서의 새로운 상태는 단일 가우스 분포를 가지게 되는데 전체 상태가 각각 2혼합 가우스 분포의 최적인 파라미터를 가지도록 재학습하게 된다. 이후, 미리 정한 상태수에 도달하도록 단계 2, 3을 반복하게 된다.

5. 분포의 변화

지금까지의 처리를 통해서 HMnet 모델의 각 상태는 2혼합 가우스 분포를 가지게 되며, 최종적으로 HMnet 모델이 각 상태마다 단일 출력확률 분포를 가지도록 HMnet 전체를 재학습하게 된다.

위에서 나타낸 것과 같이 SSS와 SSS-free 알고리즘의 차이점은 문맥방향으로의 분할방법에서 SSS 알고리즘은 학습 샘플의 문맥환경 요소마다 그룹으로 분할하여 그 그룹마다 출력우도가 높은 상태를 선택하게 된다. 이에 반하여 SSS-free 알고리즘은 각 학습 샘플마다 출력우도가 높은 상태를 선택하는 것이 큰 차이점이라고 할 수 있다.

4. 인식실험 및 결과

본 논문에서는 SSS와 SSS-free 알고리즘에 의해 작성한 HMnet 음소모델의 유효성을 확인하기 위해 ETRI 445 단어에서 학습에 참가한 3명의 두 번째 발성에 대해 Viterbi 알고리즘[5]으로 음소인식 실험을 수행하였다.

그림 3은 SSS 알고리즘으로 상태수가 50부터 최적의 상태수인 520까지 작성한 HMnet 음소모델에 대해 ETRI 445 단어의 화자 3명을 화자종속으로 음소인식 실험한 결과를 나타낸 것이다. 또, 그림 4는 SSS-free 알고리즘으로 작성한 각 상태수에 따른 모델에 대해 ETRI 445 단어의 화자 3명을 화자종속으로 음소인식 실험한 결과를 나타낸 것이다.

또한, 표 1은 SSS와 SSS-free 알고리즘으로 작성한 HMnet과 단일 HMM 음소모델을 한국어에 적용하여 화자종속과 화자독립으로 인식실험을 수행한 결과를 비교한 것이다.

그림 3과 4의 화자종속 음소인식 결과로부터 상태수가 증가할수록 인식 성능이 향상되고 있음을 확인할 수 있다. SSS 알고리즘에 의한 화자종속인 경우(그림 3) 상태수가 520일 때 평균 62.8%의 인식률을 얻었으며, SSS-free 알고리즘에 의한 화자종속인 경우(그림4) 상태수가 420일 때 평균 64.2%의 인식률을 얻었다. 이로부터 SSS-free 알고리즘으로 작성한 모델을 이용한 경우가 SSS 알고리즘에 비해

평균 1.4%의 인식률 향상을 보임을 알 수 있다. 분할된 상태수를 비교한 경우, 문맥적인 요소를 사용하지 않은 SSS-free 알고리즘의 경우가 상태수 420에서 최적 상태에 도달한 반면 문맥적인 요소를 이용한 SSS 알고리즘의 경우가 상태수 520에서 최적인 상태에 도달하여 SSS-free 알고리즘이 적은 상태수에서 최적인 상태에 도달함을 확인할 수 있었다.

또한, HMnet 음소모델을 이용한 경우, 전체적으로 상태수가 증가할수록 음소인식률이 증가하는 것을 알 수 있었다. 최적의 상태수에서의 인식률을 비교한 결과 SSS-free 알고리즘으로 작성한 음향모델의 성능이 우수함을 알 수 있었다. 표 1에서 단일 HMM을 이용한 음소인식 결과와 비교한 경우, 최대 20%이상의 인식률 향상을 가져와 이 모델의 유효성을 확인할 수 있었다.

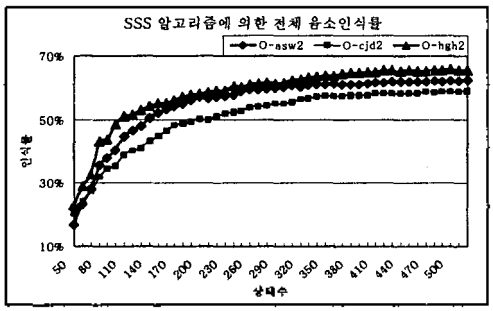

그림 3. SSS 알고리즘에 의한 화자종속 음소인식률(%)

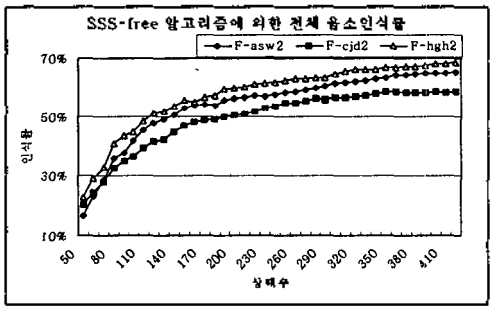

그림 4. SSS-free 알고리즘에 의한 화자종속 음소인식률(%)

표 1. HMnet과 단일 HMM 음소모델을 이용한 화자종속, 독립 음소인식률의 비교

모델 (알고리즘:상태수)	조건	화 자			평균(%)
		ASW	CJD	HGH	
HMnet (SSS:520)	종속	62.4	59.2	65.7	62.8
HMnet (SSS-free:420)	종속	65.3	58.9	68.4	64.2
		CJR	JYC	KHN	
HMnet (SSS-free:420)	독립	42.9	52.2	54.6	49.9
HMM(단일)	종속	39.7	42.9	47.7	43.4

5. 결 론

본 논문에서는 음성인식을 위한 음향모델 개량을 위한 기초적 연구로서, SSS와 SSS-free 알고리즘을 이용한 HMnet 음향모델 작성방법을 검토하고 작성한 음향모델의 유효성을 확인하기 위해 ETRI 445 단어에 대해서 화자종속 음소인식 실험을 수행하였다.

인식실험 결과, SSS 알고리즘으로 작성한 HMnet 음향모델에 대한 화자종속의 경우 상태수 520에서 평균 62.8%의 인식률을 얻었으며, SSS-free 알고리즘으로 작성한 HMnet 음향모델에 대한 화자종속의 경우 상태수 420에서 평균 64.2%의 인식률을 보였다.

이로부터 문맥적인 요소를 부가하지 않은 SSS-free 알고리즘이 SSS 알고리즘보다 적은 상태수에서 최적의 상태에 도달함을 알 수 있었으며 HMM을 이용한 음향모델에 비해 많은 인식률 향상을 보여 한국어에 HMnet 음향모델을 적용한 경우 두 알고리즘이 유효함을 확인할 수 있었다.

향후 대량의 데이터를 이용하여 인식실험을 실시하고 이 모델의 개량에 대해 검토하고자 한다.

참고문헌

[1] 김범국, 정현열, "가변장 음소모델을 이용한 음소인식," 한국음향학회지, 제16권 제8호,1997.11.

[2] K.F. Lee, S. Hayamizu, H.W Hon, C. Huang, J. Swartz, R. Weide, "Allophone Clustering for Continuous Speech Recognition," In ICASSP'90, pp. 749-752, 1990.

[3] Takami, J, Sagayama, S, "A Successive State Splitting Algorithm for Efficient Allophone Modeling," In ICASSP'92, pp. 573-576, 1992.

[4] Motoyuki Suzuki, Shozo Makino, "A New HMnet Construction Algorithm requiring No Contextual Factors," IEICE TRANS. INF. & SYST., Vol. E78-D, No. 6, 1995.

[5] L.R. Rabiner, B.H. Juang, "Fundamentals of Speech Recognition,""Prentice Hall, 1993.

[6] 임영춘, 오세진, 김범국, 정현열, "HMnet을 이용한 한국어 음소인식에 관한 연구," 한국음향학회 영남지회, 제7권 pp. 50-53, 2000. 10.