연속 음성 인식 시스템을 위한 향상된 결정 트리 기반 상태 공유

Improved Decision Tree-Based State Tying In Continuous Speech Recognition System

김 동 화*, Xintian Wu**, Chaojun Liu**, 김 형 순***, 김 영 호****
(Dong Hwa Kim*, Xintian Wu**, Chaojun Liu**, Hyung Soon Kim***, Young Ho Kim****)

요 약

결정 트리 기반 상태 공유 방법은 HMM을 사용하는 많은 연속 음성 인식 시스템에서 강인하고 정확한 문맥 종속 음향 모델링 뿐만 아니라 훈련 중에는 나타나지 않은 모델들의 합성을 위하여 널리 사용되고 있다. 음성 결정 트리를 구성하기 위한 표준적인 방법은 단일 가우시안 트라이폰 모델을 이용한 1계층 프루닝만을 사용하고 있다. 본 논문에서는 더욱 정교한 음향 모델링을 통하여 인식 성능 향상을 도모하기 위하여 새로운 2가지 접근 방법 즉, 2계층 결정 트리와 복수 혼합 결정 트리를 제안한다. 2계층 결정 트리는 상태 공유와 혼합 가중치 공유를 위하여 2계층 프루닝을 수행하며, 두 번째 계층을 사용하여 공유 상태들도 음성 문맥의 유사도에 따라서 서로 다른 가중치들을 사용할 수 있다. 두 번째 제안된 방법에서는 훈련 과정 즉, 혼합 분할 및 재추정 과정과 함께 음성 결정 트리가 계속 갱신되어진다. 복수 혼합 결정 트리를 구성하기 위하여 단일 가우시안 뿐만 아니라 복수 혼합 가우시안 모델이 함께 사용된다. 제안된 방법들을 이용하여 BN-96과 WSJ5k 데이터를 사용한 연속 음성 인식 실험을 수행한 결과, 표준 결정 트리를 사용한 시스템과 비교하여 공유 상태의 개수를 비슷하게 유지하면서 단어 오인식률을 줄일 수 있었다.

ABSTRACT

In many continuous speech recognition systems based on HMMs, decision tree-based state tying has been used for not only improving the robustness and accuracy of context dependent acoustic modeling but also synthesizing unseen models. To construct the phonetic decision tree, standard method performs one-level pruning using just single Gaussian triphone models. In this paper, two novel approaches, two-level decision tree and multi-mixture decision tree, are proposed to get better performance through more accurate acoustic modeling. Two-level decision tree performs two level pruning for the state tying and the mixture weight tying. Using the second level, the tied states can have different mixture weights based on the similarities in their phonetic contexts. In the second approach, phonetic decision tree continues to be updated with training sequence, mixture splitting and re-estimation. Multi-mixture Gaussian as well as single Gaussian models are used to construct the multi-mixture decision tree. Continuous speech recognition experiment using these approaches on BN-96 and WSJ5k data showed a reduction in word error rate comparing to the standard decision tree based system given similar number of tied states.

*밀양대학교 정보통신공학과
** Ph. D candidate, Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, USA
***부산대학교 전자공학과
****부산대학교 전자계산학과
접수일자 : 1999년 3월 22일

Ⅰ. 서 론

연속 분포 은닉 마르코프 모델(Continuous Hidden Markov Models(CHMM))에 기반한 음성 인식 시스템에서 높은 인식률을 얻기 위해서 문맥 종속 모델들과 함께 혼합(mixture) 가우시안 출력 확률 분포가 주로 사용되어지고 있다[1]. 이 경우에 많은 숫자의 모델 파라미터들로 인한 데이터 부족 문제가 필수적으로 수반된다. 또한 모델별 훈련 데이터의 수가 균일하지 않은 경우가 일반적이다. 이러한 이유로 CHMM 기반 연속 음성 인식 시스템을 위한 훈련용 데이터의 가용성과 모델의 복잡성 사이에 균형을 유지할 수 있는 방법이 제공되어야 한다[2]. 이를 위한 대표적인 접근 방법으로는 평활화(smoothing) [3][4]와 공유[5]가 있다. 평활화는 주로 보간법을 사용하여 좀더 구체적인 모델과 덜 구체적인 모델의 파라미터들을 평활화 시킨다. 이 방법은 평활화를 수행하기 전의 모델이 가졌던 문맥종속성은 유지하면서 다른 문맥들의 훈련용 데이터를 공유함으로써 모델의 강인성을 함께 유지할 수 있다. 그러나 이 방법에서는 한 모델내의 각 부분들이 다른 모델내의 여러 가지 다른 부분들과 평활화가 이루어질 수 있는 문제점이 있다[6]. 공유는 공유 수준에 따라 모델 기반 공유와 상태(state) 기반 공유로 나누어 질 수 있으나 모델 기반 공유에서는 왼쪽과 오른쪽의 문맥들을 독립적으로 처리할 수 없기 때문에 상태 기반 공유 방법이 더 우수한 것으로 알려져 있다[2]. 상태 기반 공유는 상향식(bottom-up)과 하향식(top-down) 방법으로 구현될 수 있다. 상향식 방법은 원래의 모델들을 대상으로 하여 파라미터들을 공유할 수 있는 집합들을 결정하며 이러한 집합들을 사용함으로써 보다 강인한 모델들을 추정할 수 있다. 그러나 상향식 방법은 군집화에 사용될 모델 파라미터들의 초기 추정치를 생성하기 위하여 각 문맥(context)들에 대한 데이터를 반드시 필요로 하기 때문에, 인식 중에 나타나지만 훈련 중에는 나타나지 않는 데이터들에 대한 모델을 구성할 수 없으며 단순히 back-off를 사용해야만 하는 문제점이 있다[9]. 하향식 방법으로는 결정 트리에 기반한 상태 공유[2][6][10]가 주로 사용되고 있다. 이 방법은 음향학적으로 비슷한 문맥들을 결정하기 위하여 훈련용 데이터와 함께 언어학적인 지식을 트리 구성시에 사용한다. 상향식 방법에 비하여 모든 문맥들에 대한 문맥 종속 모델을 찾을 수 있으므로 훈련용 데이터에는 나타나지 않았던 문맥들을 모델링 할 수 있으며, 질의어의 집합을 생성할 때 전문가의 지식이 사용되어질 수 있고, 트리 생성 과정에 적당한 제약을 둠으로써 각 모델별로 충분한 데이터를 확보하면서 동시에 상태들의 개수를 적정 수준으로 유지할 수 있다는 등의 장점을 가진다. 이러한 이유들로 결정 트리 상태 공유 기반 음향 모델링이 HMM을 사용한 연속 음성 인식 시스템에서의 음성 스펙트럼의 변화를 모델링하기 위한 방법으로 점점 인기를 얻어가고 있다[11][12].

위에서 기술한 표준적인 음성 결정 트리는 단일 가우시안 모델을 사용하고 1계층(level) 프루닝(pruning)만을 수행하여 생성되어진다. 1계층 프루닝 구조에서는 트리의 프루닝이 수행된 후 트리의 각 leaf는 문맥 특성이 비슷한 상태들의 집합을 나타내며 동일 집합에 포함된 모든 상태들은 하나의 가우시안을 공유하게 된다. 다음 단계인 가우시안 분할(splitting)과 재추정(re-estimation) 과정이 수행된 후 동일 노드에 포함된 모든 상태들은 훈련 및 인식 과정에서 동일한 혼합 가중치(mixture weight)를 공유하게 된다. 이러한 방법으로 생성된 트리에 기반한 상태 공유 정보를 사용할 경우 음향 모델링의 정확도가 떨어지며 이것은 시스템의 인식 성능 저하를 가져올 수 있는 한가지 요인이 될 수 있다. 본 연구에서는 보다 정확한 음향 모델링을 위하여 새로운 2가지 접근 방법 즉, 단일 가우시안 모델을 사용하는 2계층 프루닝 결정 트리와 단일 가우시안과 복수 혼합(multi-mixture) 가우시안을 함께 사용하는 복수 혼합 결정 트리를 제안한다. 2계층 트리 프루닝 구조에서 첫 번째 계층은 상태 공유층(state-tying level)으로, 두 번째 계층은 가중치 공유 층(weight-tying level)으로 사용되며 이것을 이용하여 공유 상태들도 그들의 음성 문맥의 유사 정도에 따라서 서로 다른 혼합 가중치를 사용할 수 있다. 2계층 프루닝을 위한 척도(criteria)로 1계층 프루닝에서 사용한 것과 동일한 로그 우도의 증감을 사용하였다. 두 번째 방법인 복수 혼합 결정 트리는 보다 정교하게 음성 결정 트리를 구성하기 위하여 단일 가우시안 모델을 사용한 결정 트리의 구성만으로 끝내지 않고 다음 단계인 가우시안 분할 및 재 추정 과정과 함께 음성 결정 트리도 계속 갱신되어진다. 이를 위하여 복수 혼합 결정 트리는 단일 가우시안 모델 뿐만 아니라 복수 혼합 가우시안 모델을 함께 사용하여 구축된다. 본 논문에서는 먼저 표준적인 결정 트리 기반 상태 공유 방법에 대하여 설명한 다음, 본 연구에서 제안하는 2계층 결정 트리와 복수 혼합 결정 트리에 기반한 상태 공유 알고리즘과 이들을 연속 음성 인식 시스템에 적용하는 방법 및 실험 결과에 대하여 차례로 기술한다.

Ⅱ. 결정 트리 기반 상태 공유

이 장에서는 상태 공유를 위한 표준적인 결정 트리 구성 방법에 대하여 기술한다. 음성(phonetic) 결정 트리는 각 노드에 하나씩의 음성 질의(phonetic question)를 가지고 있는 이진 트리이다. 음성 질의의 형태는 다음과 같다.

Is right Nasal (*m, *n, *en, *ng, *em) ?

Is left Central (t*, d*, en*, n*, s* ...) ?

음성 결정 트리의 모양은 그림 1과 같으며 루트 노드에는 특정 음소(phoneme)의 상태들의 풀(pool)이 입력되며

각 노드별 최적의 음성 질의들을 사용한 분할 과정을 거쳐서 마지막으로 생성된 각각의 leaf 노드는 비슷한 성질의 문맥을 가지는 상태들을 포함하게 된다.

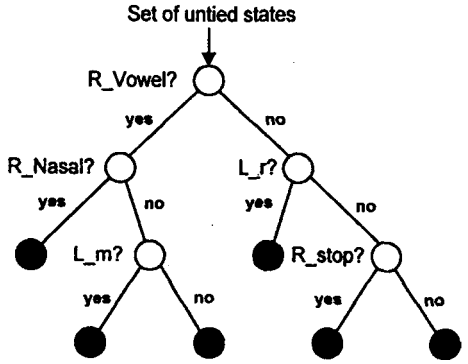

그림 1. 음성 결정 트리의 예
Fig. 1. An example of phonetic decision tree.

트리 생성을 위한 노드 분할 과정은 다음과 같은 방법으로 진행된다. 먼저 질의를 사용하여 생성된 yes/no 노드들과 부모 노드간에 로그 우도의 이득을 계산한다. 로그 우도의 이득은 식 (1)을 사용하여 계산된다.

$$\Delta L = L(A) - (L(B) + L(C)) \tag{1}$$

A는 부모 노드에 포함된 상태들의 집합이며, B와 C는 자식 노드(yes/no) 각각에 포함된 상태들을 나타낸다. L(A)는 훈련용 프레임들의 집합 F를 발생시키는 A의 로그우도이며, 다음 식 (2)를 사용하여 근사적으로 계산된다. 여기서 A에 속한 모든 상태들은 평균 μ_A와 공분산 Σ_A를 공유하며 천이 확률은 무시되는 것으로 가정한다. γ_s^f는 프레임 o_f가 상태 s에서 발생되는 확률이며, 이것은 상태 점유도(state occupancy count)라고도 불리며 Baum-Welch 재추정을 수행할 동안에 누적된다.

$$L(A) = \sum_{f \in F} \sum_{s \in A} \log(P(o_f; \mu_A, \Sigma_A)) \gamma_s^f \tag{2}$$

출력 확률 분포가 가우시안이라고 가정하면 L(A)는 식 (3)을 사용하여 계산될 수 있다. n은 특징 벡터의 차수를 나타내며, Σ_A는 훈련용 데이터로부터 직접 계산되는 것이 아니라 훈련 과정에서 생성된 상태들의 파라미터를 사용하여 계산된다.

$$L(A) = -\frac{1}{2}(\log(|\Sigma_A|) + n(1 + \log(2\pi))) \sum_{s \in A} \sum_{f \in F} \gamma_s^f \tag{3}$$

L(B)와 L(C)도 L(A)와 동일한 방법으로 계산된다. 식 (1)을 사용하여 가장 큰 로그 우도의 이득을 생성하는 질의가 그 노드의 질의로 선택되며 그것을 사용하여 부모 노드의 데이터를 yes와 no의 두 집합으로 분할한다. 가장 적합한 질의를 사용하였을 때의 이득이 특정 문턱치(threshold)보다 작아질 때까지 이러한 과정을 반복한다. 훈련성(trainability)을 위한 제약 조건으로써 γ_s에 대한 문턱치도 함께 사용된다. 분할 과정이 끝나면 트리의 각 leaf는 특정 음성의 상태들 가운데 비슷한 성질을 가진 상태들을 포함하게 된다. 그러나 종종 서로 다른 leaf 노드들이 매우 비슷한 형태가 되는 경우가 발생하며, 이것은 음성 문맥(phonetic context)들은 달라도 특징 벡터 내에서의 변화들은 같을 수 있다는 사실에 기인한다. 따라서 leaf들 가운데 비슷한 노드들이 있는지를 검사하여 이러한 노드들에 대한 합병(merge)을 수행함으로써 공유 상태들의 수를 줄일 수 있다.

Ⅲ. 제안된 결정 트리 알고리즘

3.1 2계층 결정 트리

앞에서 기술된 표준적인 결정 트리 기반 상태 공유 방법을 사용하는 시스템에서는 상태 공유 층이 결정되면 같은 노드에 포함되는 모든 상태들은 다음 단계인 가우시안 분할 과정 후에 서로 동일한 혼합 가중치들을 사용하게 된다. 이 절에서는 공유 상태(tied state)들의 개수는 그대로 유지하면서 보다 정확한 음향 모델링을 위하여 2계층 프루닝 결정 트리를 구성하는 방법에 대하여 기술한다. 두 번째 계층인 가중치 공유 층을 결정하기 위하여 표준적인 결정 트리에서 사용되고 있는 1계층 프루닝과 거의 동일한 방법이 사용된다. 2계층 프루닝 에서는 그림 2처럼 상태 공유 층 외에 가중치 공유 층이 추가로 결정되며, 두 번째 계층을 사용함으로써 동일 클러스터 내의 상태들이 서로 다른 혼합 가중치를 사용할 수 있게 된다.

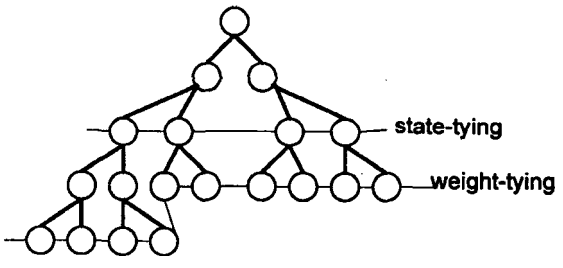

그림 2. 2 계층 결정 트리
Fig. 2. 2 level decision tree.

2 계층 결정 트리 구성을 위한 알고리즘은 그림 3과 같다. 그림 3의 단계 (1)에서는 이득을 계산하기 위하여 앞장의 식 (1)이 사용되며, 단계 (2)에서 노드 분할을 할 때

어떤 노드가 지나치게 적은 문맥(context)들을 포함하고 있게 되는 경우를 피하기 위하여 outlier 문턱치를 사용하여 분할 수행 여부를 다시 한번 결정하게 된다. 단계 (4)에

input: states pool of a phone
output: 2 level phonetic-decision tree

(1) Find a node and question that give maximum gain in log likelihood.
(2) If gain exceeds threshold then split node into yes/no children and go to (1).
(3) Through the leaf list, find 2 nodes giving minimum decrease in log likelihood when merging. If the decrease is lower than threshold then merge the second node into the first node.
(4) Compute total and splitting probability of all leaf nodes for the best questions, reset the threshold, and find a typical state(tstate) of each cluster.
(5) Do (1) and propagate tstate to split nodes.
(6) If gain exceeds new threshold then split node into yes/no children and go to (6).
(7) Through the leaf list, find 2 nodes giving minimum decrease in log likelihood when merging. If the decrease is lower than threshold and the 2 nodes have same tstate then do merge else do not merge.

그림 3. 2 계층 결정 트리 알고리즘
Fig. 3. Algorithm of two-level decision tree.

서는 2계층 프루닝을 위하여 각 leaf 노드에 대하여 전체 확률과 분할 확률을 계산하고, 각 상태 집합에 대한 가장 대표적인 상태를 선택하여 해당 leaf 노드에 기록한다. 단계 (5)에서 2단계 분할을 수행할 때 이 상태들은 후손 노드들에게 전파된다. 분할 과정이 끝난 후 마찬가지로 비슷한 집합들에 대한 합병이 수행되며, 이 때 두 노드가 서로 다른 상태를 가지고 있으면 합병을 수행하지 않고, 동일한 상태를 가지는 leaf 노드들만을 대상으로 하여 합병을 수행한다. 본 실험에서는 분할과 합병을 위하여 동일한 문턱치를 사용하였다. 가중치 공유 층이 결정되면 이것을 사용하여 가중치와 공유 상태들을 위한 원형(template)을 생성한다. 이 원형은 훈련 과정에서 가우시안 혼합 분할 후에 생성되는 새로운 가중치 집합들을 위해서 사용된다. 또한 상태 공유 과정에서 새롭게 생성된 물리적(physical) 모델과 기존의 논리적(logical) 모델간에 매핑을 위한 방법과 인덱스 구조 등을 구현하였으며, 이를 위하여 가중치와 혼합 집합에 대한 포인터들을 필드로 가지는 확장 상태 집합이라는 자료 구조를 사용하였다.

3.2 복수 혼합 결정 트리

이 절에서는 보다 정확한 음향 모델링을 위하여 본 연구에서 제안하는 두 번째 방법인 복수 혼합 결정 트리 알고리즘에 대하여 기술한다. 이 방법에서는 단일 가우시안 모델 뿐만 아니라 복수 혼합 가우시안 모델을 함께 사용하여 상태 공유를 위한 음성 결정 트리가 구성 및 갱신된다. 최초로 생성되는 단일 가우시안 결정 트리는 노드 분할을 위하여 표준적인 방법과 동일한 로그 확률값의 근사치를 사용한다. 이 결정 트리를 사용하여 확보되는 적당한 양의 데이터로써 재 추정과 혼합 분할 과정이 수행된다. 이러한 훈련 과정 즉, 재추정 및 혼합 분할이 수행된 이후에 복수 혼합 가우시안 모델을 사용하는 결정 트리가 다시 생성된다. 음향 모델링을 위한 이러한 일련의 과정은 혼합 구성 요소(component)들이 특정 개수에 도달하거나 시스템의 성능이 원하는 수준이 될 때까지 반복된다.

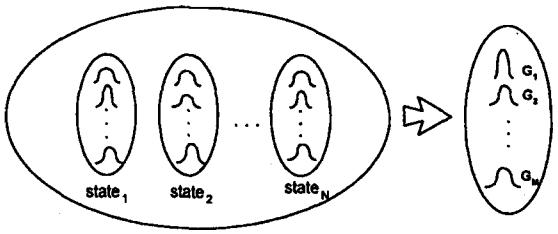

그림 4. 복수 혼합 상태들의 풀에 대한 파라미터 계산
Fig. 4. Pooled parameter for the pool of multi-mixture states.

복수 혼합 결정 트리를 생성하는 과정에 각 노드에 대한 확률의 근사 값이 구해져야 하며, 이것을 위하여 먼저 복수 혼합 상태들의 풀(pool)에 대한 파라미터들 즉, 평균, 분산 및 총 상태 점유도가 계산되어야 한다. 복수 혼합 상태들의 풀의 모양은 그림 4에 나타내었고 상태 풀에 대한 파라미터들은 다음 식 (3),(4),(5)를 사용하여 계산된다.

$$\gamma_j = \sum_{i=0}^{N} \gamma_{ij}, \quad 1 \le j \ge M, 1 \le i \ge N \tag{3}$$

$$\mu_j = \frac{\sum_{i=1}^{N} \gamma_{ij}\mu_{ij}}{\gamma_j} \tag{4}$$

$$\Sigma_j = \frac{\sum_{i=1}^{N} \{\gamma_{ij}(\mu_{ij} - \mu_j)^2 + \Sigma_{ij}\}}{\gamma_j} \tag{5}$$

여기서,

N : 상태 풀에 포함되는 상태들의 개수
M : 상태별 혼합 구성 요소들의 최대 개수
$\gamma_{ij}, \mu_{ij}, \Sigma_{ij}$: i번째 상태에서의 j번째 가우시안의 상태 점유 확률, 평균 및 분산
$\gamma_j, \mu_j, \Sigma_j$: 상태 풀의 j번째 가우시안의 상태 점유 확률, 평균 및 분산

을 나타낸다. 상태 풀에 대한 확률의 근사 값을 계산하기 위하여 상태들에 대한 데이터의 할당은 결정 트리 생성 중에는 변하지 않으며, 상태 풀에 데이터의 전체 우도는 상태 풀의 파라미터를 사용하여 혼합별 상태 점유 확률이 가중된 각 혼합의 로그 우도의 단순 평균으로써 근사적으로 계산될 수 있다고 가정한다. 이러한 가정 아래 데이터의 로그 우도 근사값 LL 은 다음 식을 사용하여 계산된다.

$$LL = \sum_{o_t:s_t \in S} \sum_{i=1}^{M} \log\{P(o_t;G_i)w_i\}\gamma_i$$
$$= \sum_{o_t:s_t \in S} \sum_{i=1}^{M} \left\{ -\frac{1}{2}\{n\log(2\pi) + \log(|\Sigma_i|) + (o_t-\mu_i)'\Sigma_i^{-1}(o_t-\mu_i)\} + \log(w_i) \right\}\gamma_i \quad (6)$$

여기서,

n : 특징 벡터의 차수

S : 상태들의 클러스터

o_t : 특정 상태 정렬(alignment) s에 할당된 모든 프레임

G_i : 혼합의 i번째 가우시안

w_i : i번째 가우시안에 대한 가중치

를 의미한다. 위의 LL 값을 데이터를 사용하여 직접 계산할 경우 메모리와 계산량 면에서 비용이 매우 높아지므로 각 상태들이 이미 가지고 있는 통계량을 사용하여 LL 값을 계산한다. 이것을 위하여 식 (6)에 가우시안 재추정식을 대입하면 근사 식 (7)을 얻을 수 있다.

$$LL = \sum_{i=1}^{M} \left\{ -\frac{1}{2} \sum_{j=1}^{n} (1 + \log(2\pi) + \log(\Sigma_{ij})) + \log w_i \right\} \gamma_i \quad (7)$$

여기서,

n : 특징 벡터의 차수

Σ_{ij} : 공분산 의 j번째 대각 원소

를 각각 나타낸다. 결정 트리 구성을 위한 노드 분할 과정은 전체 노드와 분할된 노드들에 대한 위의 LL값에 의해 제어된다. 분할 과정이 끝난 후 leaf 노드들에 대한 합병이 수행되며 이 때도 역시 LL값이 사용된다. 전체적인 음향 모델링 과정에서 결정 트리의 leaf들의 개수는 혼합 구성 요소들의 숫자가 증가함에 따라 점차 감소시킨다. 또한 이 접근 방법에서도 모든 음소에 대하여 각 상태별로 1개씩의 결정 트리가 만들어진다.

Ⅳ. 실험 및 결과

4.1 시스템 구성 및 구현

본 연구에서 제안한 방법들에 대한 실험을 위하여 2개의 연속 음성 인식 시스템을 구현하였으며 이들 시스템의 사양을 표 1에 나타내었다. 먼저 HUB4는 DARPA/NIST의 대어휘 연속 음성 인식 기술 평가를 위하여 구축된 시스템이다. 여기서 사용된 BN-96 데이터는 HUB4-1996을 위하여 LDC에서 수집하여 배포한 방송 뉴스의 음성 및 텍스트 코퍼스(corpus)이다. 이 데이터에는 대역별, 배경 잡음 정도, 배경 음악 유무, 성별 등에서 서로 다른 음성데이터 들이 포함되어 있으며 장르별로 쇼, 애피소드, 이야기로 구분되어 있다. 이 시스템에서 사용된 triphone은 3개의

표 1. HUB4 및 WSJ 시스템별 사양
Table 1. HUB4 and WSJ system specification.

사양 \ 시스템	HUB4	WSJ
HMM	within/cross-word triphone	cross-word triphone
특징 벡터	39차(MFCC,CMS)	39차(MFCC)
어휘 수	56k	5k
최대 혼합 수	16/24	10
언어 모델	back-off trigram,bigram	back-off bigram
디코더	2 pass	1 pass
적응화	MLLR	none
음성 질의의 개수	217	217
훈련용 데이터	BN-96 data(50 hours)	SI84 data, 7200 sent.s
평가용 데이터	BN-96 data(1000 sec)	si-dt-05 set, 442 sent.s

상태들과 각 상태별 최대 24개의 가우시안들을 가지는 CHMM으로써 모델링된다. 디코딩은 2단계로 수행되며 1단계에서는 bigram과 within-word triphone 모델을 사용하여 상당히 시간이 많이 소요되는 bigram 그래프가 생성되며 탐색을 위한 음향 모델을 구성하기 위하여 평면(flat) 구조가 아닌 사전 트리 구조가 사용되었다. 2단계는 거의 실시간으로 실행되며 여기서는 cross-word triphone 모델을 사용하여 trigram 그래프 탐색과 단어 기반 N-Best가 수행된다. 디코딩의 각 단계에서 적응화를 위해서 Maximum Likelihood Linear Regression(MLLR)이 사용되었고 Bayesian information criteria(BIC)를 디코딩 단계의 세그멘테이션에서 사용하였다. WSJ는 Wall Street Journal 데이터를 훈련 및 평가용으로 사용하는 시스템으로서 본 실험에서는 WSJ 5k 데이터를 사용하였다. 이 시스템에서는 HMM 상태별 최대 10개의 가우시안들과 1단계 디코더가 사용되었다.

본 연구를 위하여 28개의 pentium CPU를 가지는 9대의 PC들이 사용되었다. 훈련, 디코딩, 결정 트리 구성 및 언어 모델링을 위하여 하나의 작업을 분리 수행 가능한 단위들로 구분한 다음 사용 가능한 CPU에 할당하여 병렬로 처리하는 방법을 사용하였다. 이를 위하여 모든 머신들이 공유할 작업 목록 파일을 생성하였으며, 각 작업을 가용한 CPU에 할당하고 동기화를 담당할 스케줄러 프로그램이 사용되었다. 결정 트리 생성에서는 각 음소에 대한 상태별 결정 트리 즉, 약 135개의 트리 생성이 병렬로 처리됨으로써 트리 생성 시간을 10배 이상 단축시킬 수 있었다. 본 실험을 위하여 사용된 모든 시스템에서 HMM 모델을 효율적으로 관리하기 위하여 물리(physical), 파라미터, 매핑(mapping)의 3클래스들로 구분하여 구현하였다. 효율적이고 확장 가능한 질의 매칭을 위하여 단순 문자열 매칭보다는 음소들의 인덱스에 기반한 마스킹 방법을 사용하였으며, 훈련 과정에서 매우 자주 발생하는 새로운 HMM의 삽입을 위하여 나머지 연산 해쉬 함수가 사용되었다.

4.2 실험 결과

결정 트리를 구축하기 전 단계에서의 HMM 상태들의 개수는 HUB4 에서는 약 27500개, WSJ에서는 약 16000개 정도이었다. 공유 상태들의 개수를 결정하기 위하여 공유 상태들의 개수가 시스템의 성능에 미치는 영향에 대한 실험을 수행하였으며, 그 결과를 토대로 일반적으로 많이 사용되고 있는 HUB4는 약 7000, WSJ는 약 4000을 공유상태들의 개수로 사용하였다. 이런 경우 결정 트리 생성 이전의 상태 수의 25% 수준으로 축소시키는 결과가 된다.

먼저 2 계층 결정 트리를 사용하여 상태 군집화를 수행할 때 문턱치별 공유 상태들과 공유 가중치 집합 개수의 관계를 파악하기 위하여 실험한 결과를 표 2와 3에 나타내었다. 여기서 사용된 outlier 문턱 값은 50이며, 표에서 공유 가중치 집합들의 개수는 전체 135개 트리들의 leaf 노드들을 합한 개수와 같다.

표 2. HUB4에서 7220 공유 상태 집합에 대한 문턱치별 공유 가중치들의 개수

Table 2. The number of tied-weights for 7220 tied-states in HUB4.

가중치 공유를 위한 문턱치	가중치 집합들의 개수
200	17,000
250	16,200
300	14,700
350*	14,100
400	12,400
450	10,600

2계층 결정 트리를 사용하여 가중치 공유 집합을 결정할 때 이 집합의 개수가 공유 상태 개수의 2배를 초과하지 않으면서 2배에 근접하도록 하였다. 이것은 시스템의 전체 파라미터 개수를 많이 증가시키지 않으면서 가중치 공유의 효과를 최대한 살리기 위한 것이다. HUB4를 위해서는 14,100개의 가중치 공유 집합들이 사용되고 2계층 결정 트리의 평균 leaf 노드의 개수는 약 105개이다. WSJ에서는 7,900개의 가중치 공유 집합들이 사용되었고 이 시스템의 2계층 결정 트리의 평균 노드 개수는 약 58개 이다. HUB4의 7220 공유 상태 집합을 위해서 문턱치로 750이 사용되고, WSJ에서 4050 공유 상태 집합을 위해서는 문턱치 400이 사용되었다.

표 3. WSJ에서 4050 공유 상태 집합에 대한 문턱치별 공유 가중치들의 개수

Table 3. The number of tied-states for 4050 tied-states in WSJ.

가중치 공유를 위한 문턱치	가중치 집합들의 개수
70	9,300
100	8,700
130*	7,900
200	6,100
250	4,700

복수 혼합 결정 트리는 훈련 과정에서 혼합 수가 점점 증가할수록 leaf들의 개수는 감소되어 마지막 단계에서 원하는 공유 상태들의 개수가 유지될 수 있도록 하였다. 단일 가우시안 결정 트리로부터 출발하여 2혼합, 4혼합 그리고 8혼합 결정 트리까지를 사용하여 인식 실험을 수행하였으며 그 결과 2혼합 결정 트리를 사용하였을 경우 인식 성능이 상당히 향상된 반면에, 4혼합과 8혼합 결정 트리는 표준 결정 트리와 비슷한 성능을 보여 주었다. 이것은 상태 pdf에 대한 로그 likelihood를 계산하기 위한 근사식 사용에 따른 오차와 시스템 튜닝 부족에서 기인하는 것으로 판단된다.

표 4. 2 혼합 결정 트리의 공유 상태 개수

Table 4. The number of tied-states of 2 mixture decision tree.

HUB4		WSJ	
1 mix. tree 공유 상태 수(문턱값)	2 mix. tree 공유 상태 수(문턱값)	1 mix. tree 공유 상태 수(문턱값)	2 mix. tree 공유 상태 수(문턱값)
8983(500)	5144(1750)	3820(360)	2990(1150)
9062(480)	5993(1650)	4310(310)	3211(1000)
10520(450)	6522(1550)	4770(290)	3754(900)
12050(400)	7082(1500) *	5010(240)	4108(750) *
13840(360)	7421(1400)		

표 4에서는 2혼합 결정트리에 대한 문턱 값과 공유 상태들의 개수를 나타내었다. 단일 가우시안 결정 트리의 공유 상태들의 개수를 결정하기 위하여 사전 실험을 하였으며, 그 결과 2혼합 결정 트리의 공유 상태 개수의 약 1.5에서 2배 정도가 적합함을 알 수 있었다. 2혼합 결정 트리의 공유 상태 개수는 2계층 및 표준 결정 트리와 마찬가지로 HUB4에서는 약 7000개, WSJ에서는 약 4000개를 유지하였다. 각 시스템을 사용하여 연속 음성 인식 실험을 수행한 결과로 얻은 단어 오인식률을 표 5에 나타내었다. 표의 2계층 결정 트리에서 괄호안의 숫자는 공유 가중치들의 개수를 의미한다. 2계층 결정 트리와 복수 혼합 결정 트리 모두 표준 결정 트리보다 단어 오인식률이 낮음을 알 수 있었다. 2계층 결정 트리의 경우 표준 트리에 비하여 깊이가 기껏해야 1 증가하게 된다. 그러나 디코딩에서 결정 트리 매핑이 전체 수행 시간에서 차지하는 비중은

표 5. 단어 오인식률 비교

Table 5. Comparison of word error rates.

	표준 결정 트리		2 계층 결정 트리		복수 혼합 결정 트리	
	HUB4	WSJ	HUB4	WSJ	HUB4	WSJ
공유 상태 개수	7220	4050	7220 (14100)	4050 (7900)	7082	4108
단어 오차율(%)	29.4	11.2	27.3	8.8	27.0	9.2

매우 적기 때문에 이 정도의 깊이 증가로 인한 수행 시간 오버헤드는 충분히 무시할 수 있다.

본 실험을 통하여 연속 음성 인식 시스템의 음향 모델링을 위하여 2혼합 결정 트리와 2계층 결정트리 모두 표준적인 방법보다는 더 나은 성능을 보여주었고, 인식 성능 향상률은 비슷하지만 트리 구성 방법이 간단한 2계층 결정트리가 더욱 우수한 접근 방법임을 알 수 있었다.

V. 결 론

본 논문에서는 보다 정확한 음향 모델링을 통하여 연속 음성인식 시스템의 인식률을 향상시키기 위하여 2계층 결정 트리 및 복수 혼합 결정 트리 기반 상태 공유 방법을 제안하였다. 표준적인 결정 트리 기반 상태 공유에서는 단일 가우시안의 문맥 정보를 이용하여 1계층 즉, 상태 공유 층이 결정되면 공유 상태들은 모두 동일한 혼합 가중치를 가진다. 반면 2계층 결정 트리에서는 새로운 가중치 공유 층을 설정함으로써 공유 상태들의 개수는 동일하게 유지하면서 공유 상태들이 음성 문맥의 유사도에 따라서 서로 다른 가중치를 사용할 수 있다. 복수 혼합 결정 트리에서는 단일 가우시안뿐만 아니라 복수 혼합 가우시안 모델을 함께 사용하여 혼합 분할 및 재 추정 과정과 함께 음성 결정 트리가 다시 생성된다. 본 연구에서 제안된 방법들을 이용하여 BN-96과 WSJ5k 데이터를 사용한 연속 음성 인식 실험을 수행한 결과 표준 결정 트리를 사용한 시스템과 비교하여 공유 상태의 개수를 비슷하게 유지하면서 단어 오차율을 2% 이상 줄일 수 있었으며, 두 방법 중 트리 구성 방법이 간단한 2계층 결정트리가 더욱 우수한 접근 방법임을 알 수 있었다.

앞으로 보다 향상된 결정 트리를 구성하기 위하여 지금 사용되고 있는 상태 위치 제한 조건을 없애고 한 음소의 모든 상태들을 대상으로 군집화를 수행하며, 질의어를 좀더 복합적인 형태로 신중하게 구성할 필요가 있다고 사료된다. 문맥 종속성을 더 잘 모델링하기 위하여 퀸폰(quinphone) 등이 사용되어야 할 것이며, 이 경우 질의어를 효율적으로 확장하는 방법이 강구되어야 할 것이다. 또한 결정 트리가 훈련 데이터를 더욱 정확하게 표현할 수 있도록 cross validation과 같은 기법을 사용할 필요가 있다.

참 고 문 헌

1. C. H. Lee, F. K. Soong and K. K. Paliwai, Automatic Speech And Speaker Recognition, Kluwer Academic Publishers, 1996.
2. S. J. Young, J. J. Odell and P. C. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modelling," Proceedings ARPA Workshop on Human Language Technology, Merrill Lynch, pp. 286-291, 1994.
3. F. Jelinek and R. L. Mercer, Pattern Recognition in Practice, North-Holland, pp. 381-397, 1980.
4. K. F. Lee and H. W. Hon, "Speaker Independent Phone Recognition Using Hidden Markov Models," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 37, No. 11, pp. 1641-1648, 1989.
5. S. J. Young, "The General Use of Tying in Phoneme-Based HMM Speech Recognisers," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, San Fransisco, pp. 569-572, 1992.
6. J. J. Odell, The Use of Context in Large Vocabulary Speech Recognition, Ph. D Thesis, Cambridge, 1995.
7. K. F. Lee, "Context-Dependent Phonetic Hidden Markov Models for Speaker Independent Continuous Speech Recognition," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 38, No. 4, pp. 599-609, 1990.
8. M. Y. Hwang, X. Huang and F. Alleva, "Predicting Unseen Triphone with Senones," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Minneapolis, pp. 311-314, 1993.
9. M. Y. Hwang, Subphonetic Acoustic Modeling for Speaker Independent Continuous Speech Recognition, Ph. D Thesis, CMU, 1993.
10. L. R. Bahl, P. V. de Souze, P. S. Gopalakrishnan, D. Nahamoo and M. A. Picheny, "Decision Trees for Phonological Rules in Continuous Speech," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Toronto, pp. 185-188, 1991.
11. W. Reichl and W. Chou, "Decision Tree State Tying Based on Segmental Clustering For Acoustic Modeling," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Seattle, pp. 801-804, 1998.
12. K. Beulen and H. Ney, "Automatic Question Generation for Decision Tree Based State Tying," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Seattle, pp. 805-808, 1998.

▲김 동 화(Dong-Hwa Kim)
현재: 밀양대학교 정보통신공학과 부교수
한국음향학회지 제 16권 6호 참조

▲Xintian Wu
1993. 9 Electrical Engineering, Tsinghwa Univ. (BS)
1996. 7 Electrical Engineering, Tsinghwa Univ. (MS)
1999.5 현재 : Ph. D candidate, Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, USA

▲Chaojun Liu
1994. 7 Univ. of Science and Technology of China(BS)
1997. 7 Institute of Acoustics, CAS (MS)
1999. 5 현재 : Ph. D candidiate, Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, USA

▲김 형 순(Hyung-Soon Kim)
현재 부산대학교 전자공학과 부교수
한국음향학회지 제 16권 5호 참조

▲김 영 호(Young-Ho Kim)
현재 부산대학교 전자계산학과 부교수
한국음향학회지 제 16권 6호 참조