

결정트리 기반 상태공유 모델 최적화에 관한 연구

한명희, 이호준, 김순협
광운대학교 컴퓨터공학과

A Study on Optimization of Decision Tree based State Tying Model

Myung-Hee Han, Ho-Jun Lee, Soon-Hyob Kim
Dept. of Computer Engineering, Kwangwoon University

요약

본 논문에서는 공유 모델링의 대표적인 방법인 결정트리 기반 상태공유 모델을 기반으로 하여 그 출력 확률 분포의 혼합 가우시안 수를 줄임으로써 모델을 최적화하고자 하였다. 결정트리 기반의 상태공유 모델링은 일반적인 방법을 따랐으며 혼합 가우시안 수를 늘려 인식률이 최대가 되는 지점에서 혼합 가우시안을 클러스터링하여 그 수를 줄였다. 클러스터링 시에 필요한 거리 측정 방법이나 가까운 두 가우시안의 합성 방법을 여러 기법을 실험하였다. 이때 인식률은 클러스터링 이전인 97.2%를 유지하였으며 총 혼합 가우시안의 감소율은 1.0%를 보임으로써 모델을 최적화할 수 있었다.

1. 서론

은닉 마코프 모델(hidden Markov model)은 음성의 변화를 모델링 하는데 효과적인 방법으로 널리 쓰이고 있다. 실제 음성의 변화를 정교하게 모델에 반영하기 위해서는 많은 수의 모델과 그것과 관련된 출력 확률 분포가 필요하게 된다.[1] 이때 출력 확률 분포가 연속 밀도인 경우 좀 더 세밀한 모델링이 가능해지지만 그 만큼 많은 데이터가 필요하게 된다. 연속 밀도 HMM에 기반을 둔 문맥 종속적인 음향 모델을 만들기 위해 가장 핵심적인 사항은 요구된 모델의 복잡도와 이용 가능한 훈련 데이터로부터 강건하게 추정될 수 있는 파라미터의 수와의 균형을 유지하는데 있다.[2]

모델들이나 모델 일부들의 파라미터 공유를 가능하게 하는 tying은 두 가지 이점을 제공한다. 첫째, 강건한 파라미터 추정에 좀 더 적은 훈련 데이터를 필요로 한다. 둘째, 메모리 사용량이나 계산 시간을 감소시킬 수 있다. 이러한 점들은 많은 양의 문맥 종속적인 모델을 만들 때 중요한 척도가 될 수 있다.[3] 결정트리를 기반으로 하는 상태공유 기법은 여러 가지 이점으로 인해 많이 쓰인다. 이 방법은 목표로 하

는 언어의 음성학적인 지식을 최대 확률 기법과 상반되지 않도록 모델에 효과적으로 반영시킬 수 있다. 음향 모델링에서 결정트리의 확률적인 기법은 규칙이나 상향식 기법에 비해 두 가지 큰 장점을 갖는다. 첫째, 결정트리의 분류나 예측 성능은 훈련 데이터에서 나타나지 않은 모델 단위나 문맥을 합성할 수 있게 해준다. 둘째, 결정트리를 기반으로 하는 상태공유의 노드 분할은 모델 선택 과정이다. 이것은 모델의 복잡성과 제한된 훈련 데이터에서 많은 수의 파라미터를 강건하게 추정하는 것 사이의 균형을 유지할 수 있는 방법을 제공한다.[4]

본 논문에서는 공유 기법에 있어서 대표적인 방법을 이용한 결정트리를 기반으로 하는 상태공유 모델을 기본 모델로 하고 그 모델의 출력 확률 분포의 혼합 가우시안 수를 거리 측정 기법을 이용하여 줄이고자 한다. 이로써 인식률은 유지하면서 메모리 사용을 줄일 수 있게 되므로 최적화된 모델을 얻게 된다. 2장에서는 결정트리 기반 상태공유 모델에 대해 설명하고 3장에서 혼합 가우시안 클러스터링의 필요성과 그 방법들에 대해 설명할 것이다. 4장은 실험을 통해 제안한 방법의 타당성을 살펴보고 5장에서 실험에 대한 결론을 볼 것이다.

2. 결정트리 기반 상태공유 모델

◆ 상태공유 모델

서론에서도 밝혔듯이 상태공유 모델의 목적은 각 음소의 문맥 종속적인 성질을 잘 나타내면서 각 상태의 출력 분포 파라미터를 강건하게 추정하기 위해 충분한 훈련 데이터를 확보하는 것을 가능하게 하는 데 있다.

이와 관련한 대표적인 생성 방법은 다음과 같다.[1]

- (1) 단일 가우시안 출력 확률 밀도 함수를 갖는 3상태 좌우 monophone 모델의 초기 집합을 구성하고 훈련한다.
- (2) 이때 이러한 monophone들의 상태 출력 분포는 공유되지 않은 문맥 종속적인 triphone 모델을 초기화하기 위해 복사되며 Baum-Welch 재추정을 사용하여 훈련한다. 전이 확률 행렬은 복사되지 않으며 각 음소의 모든 triphone이 공유하게 된다.
- (3) 동일한 monophone으로부터 유도된 triphone들의 각 집합과 대응되는 상태들이 묶이게 된다. 이런 식으로 생성된 각 클러스터에서 전형적인 상태가 본보기로써 선택되며 모든 클러스터 부분들은 이 상태에 묶이게 된다.
- (4) 각 상태에 있는 혼합 요소들의 수를 증가시키고 인식률이 최대가 되는 지점이나 원하는 수를 얻을 때까지 모델 재추정을 반복한다.

◆ 결정트리 기반 클러스터링

상향식 방법인 집접식(agglomerative)보다 결정트리 기반 클러스터링을 사용하는 이유는 훈련 데이터에 포함되지 않은 triphone을 생성해 낼 수 있기 때문이다. 이것은 응용 측면에서 본다면 가변언어회가 필요한 받아쓰기 시스템에서 사용될 수 있으며 클러스터링하는 데 시간이 오래 걸리지 않기 때문에 유용하다고 할 수 있다.

보통 결정트리는 트리의 루트 노드에서 출발하여 하향식으로 순차적인 최적화 과정을 사용하여 만들어진다. 각 노드는 훈련 데이터에서 likelihood를 최대로 증가시키는 음성학적 질의어에 따라 분할된다.[4]

S 를 HMM 상태들의 집합이라 하고 $L(S)$ 를 S 의 log likelihood라고 하자.[1] 이때 집합 S 에 있는 모든 상태들은 묶였다는 가정 하에 훈련 데이터의 프레임들의 집합인 F 에서 생성된다고 하자. 즉 공통 평균 (S)과 분산 (S)를 공유한다. 단 전이 확률은 무시

된다. 묶인 상태들의 상태 별 프레임의 정렬이 바뀌지 않는다고 가정하면 $L(S)$ 에 대해 다음과 같은 근사식을 쓸 수 있다.

$$L(S) = \sum_{f \in F} \sum_{s \in S} \log(P_{\pi}(o_f | \mu(s), \Sigma(s))) \gamma_s(o_f) \quad (1)$$

$\gamma_s(o_f)$ 는 상태 s 에 의해 생성되는 관측 프레임 o_f 의 사후 확률이다. 만일 출력 확률 밀도 함수가 가우시안이면 식(2)와 같이 쓸 수 있다.

$$L(S) = -\frac{1}{2} (\log[(2\pi)^n |\Sigma(s)|] + n \sum_{s \in S} \sum_{f \in F} \gamma_s(o_f)) \quad (2)$$

n 은 데이터의 차수를 나타낸다. 그러므로 전체 데이터 집합의 log likelihood는 상태들의 분산 $\Sigma(s)$ 와 전체 상태 점유인 $\sum_{s \in S} \sum_{f \in F} \gamma_s(o_f)$ 만으로 계산할 수 있게 된다. 상태들의 평균과 분산으로부터 앞부분을 계산할 수 있으며 상태 점유 수는 이전 단계의 Baum-Welch 재추정하는 동안 저장될 수 있다. 질의 q 에 의해 두 하위 집합 $S_y(q)$ 와 $S_n(q)$ 를 나누는 상태들 S 와 함께 주어진 노드인 경우, 그 노드는 다음 식(3)을 최대화하는 질의 q^* 를 사용하여 분할된다.

$$\Delta L_q = L(S_y(q)) + L(S_n(q)) - L(S) \quad (3)$$

식(3)은 ΔL_q 와 특정 문턱치를 넘는 $S_y(q^*)$ 과 $S_n(q^*)$ 에 대한 전체 상태 점유 수 모두를 제공한다.

3. 혼합 가우시안 클러스터링

◆ 필요성

일반적으로 전체적으로 최적화된 결정트리의 생성은 계산적으로 다룰 수 없는 문제이다. 결정트리 노드 분할에서 사용된 분포들의 파라미터 형태는 종종 가우시안 분포를 기반으로 하고 있으며 정밀한 다중 혼합 가우시안 분포는 마지막 음향 모델에서 쓰이게 된다. 이러한 불균형은 결정트리 클러스터링 과정에서 계산적인 복잡도 때문이다. 각 트리 노드의 다중 혼합 가우시안 분포는 데이터로부터 재추정될 필요가 있으며 반면에 단일 가우시안 분포는 훈련 데이터로 돌아가지 않고 클러스터들로부터 유도될 수 있기 때문이다.[4] 따라서 본 논문에서는 최종 모델인 다중 혼합 가우시안 분포들을 다시 한번 클러스터링하여 최적화하고자 하였다. 이때 최적화되었다 함은 인식률을 유지하면서 전체 혼합 가우시안의 수를 줄여 파라미터 수를 감소시키는 것을 말한다.

◆ 거리 측정

거리 측정의 목적은 가장 비슷한 것을 찾아내는 것을 의미한다. 출력 확률 분포를 연속 확률 밀도로 갖는 가우시안의 경우, Euclidean 거리법[5]과 오류율 측정에 기반을 두고 있는 가우시안 분포의 유사성을 표현하는 Bhattacharyya 거리법[6]이 있다. 본 논문에서는 두 거리 측정법을 비교하여 어느 방법이 더 나은 성능을 갖는지 확인할 것이다.

[5]에서 쓰고 있는 거리법은 다음 식(4)와 같다.

$$d(i, j) = \left[\frac{1}{n} \sum_{k=1}^n \frac{(\mu_{ik} - \mu_{jk})^2}{\sigma_{ik}\sigma_{jk}} \right]^{\frac{1}{2}} \quad (4)$$

n 은 데이터의 차수를 나타내며 μ_{ik} 와 σ_{ik} 는 상태 s (i 혹은 j)의 가우시안 분포의 k 번째 평균과 분산이다.

두 분포간의 유사성을 측정하는 일반적인 방법이 바로 Bhattacharyya 거리법이다.

$$\begin{aligned} \mu &= \frac{1}{8} (M_2 - M_1)^T \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (M_2 - M_1) \\ &\quad + \frac{1}{2} \ln \frac{|\frac{\Sigma_1 + \Sigma_2}{2}|}{\sqrt{|\Sigma_1|} \sqrt{|\Sigma_2|}} \end{aligned} \quad (5)$$

M_1 과 M_2 은 각 분포의 평균이며 Σ_1 과 Σ_2 은 공분산이다. 만일 두 분포가 정규적이라면 Bhattacharyya 거리는 Bayes 오류의 상위 범위가 된다. 이 거리법은 두 분포의 평균과 공분산이 각각 얼마나 이동했는지에 대한 차이를 측정한다.[6]

◆ 가우시안 합성

모든 가우시안 분포 간의 거리를 구한 뒤, 그 중 거리가 가장 짧은 두 분포들은 하나의 새로운 가우시안으로 형성되게끔 합쳐져야 한다. 새로운 가우시안의 파라미터들은 두 가우시안들의 파라미터로부터 유도된다.[7] 새로운 가우시안은 가중치는 w_{new} 이며 이것은 두 가우시안의 가중치의 합과 같다.

$$w_{new} = w_1 + w_2 \quad (6)$$

새 가우시안의 각 차원의 평균은 두 가우시안의 평균의 가중치가 부여된 합이며 가중치의 합에 의해 정규화 된다.

$$\mu_{new} = \frac{w_1\mu_1 + w_2\mu_2}{w_1 + w_2} \quad (7)$$

그리고 각 차원을 위한 새로운 분산 σ_{new} 은 수정된 분산의 평균의 가중치된 합이다. σ_1 과 σ_2 를 두 가우시안의 공분산이라 하고 σ_{new} 를 식(7)과 같이 정의하면

새로운 분산 σ_{new} 은 다음과 같다.

$$\sigma_{new} = w_1(\sigma_1 + (\mu_1 - \mu_{new})^2) + w_2(\sigma_2 + (\mu_2 - \mu_{new})^2) \quad (8)$$

◆ 가우시안 클러스터링

앞 절에서 언급한 거리 측정 방법과 가우시안 합성을 이용하여 클러스터링 하는 과정을 그림1에 나타내었다.

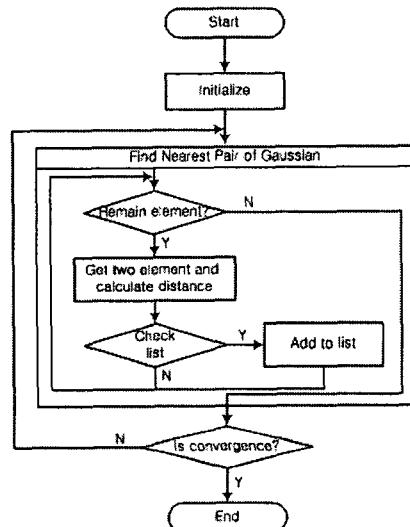


그림 1 클러스터링 과정

먼저 초기화를 통해 혼합 가우시안을 로딩한다. 가장 가까운 거리의 가우시안 분포를 찾기 위해 모든 가우시안 분포 사이의 거리를 구하고 두 개를 합성하여 새로운 가우시안을 혼합 가우시안 집합에 추가한다. 이때 리스트를 유지하는 이유는 어떤 가우시안들이 합쳐졌는지를 살펴보기 위해서이다.

4. 실험 및 결과

◆ 훈련 데이터베이스 및 모델 생성

본 논문에서 대상으로 하고 있는 음성 데이터베이스는 증권명 1680개이며 이것은 각 560개씩 3 set로 구성되어 있다. 전체 127명의 음성 데이터베이스 중 모델을 구성하는 데 116명, 테스트는 11명 분량으로 하였다. 이 데이터베이스는 8kHz, 16bit이며 preemphasis 계수는 0.97이다. Hamming 창을 25ms로 10ms 씩 이동하여 MFCC 12차와 정규화 된 에너지, 이것들의 엘타 값으로 26차의 계수를 사용하였다. 모델 훈련은 데이터베이스를 Viterbi decoding 기법을

이용하여 정렬한 데이터를 훈련 모델의 기반으로 하였으며 triphone 모델을 기준 모델로 잡았다. 결정트리를 이용한 상태공유 모델 생성 시, 질의어들은 [8]에서 정리한 한국어 변이음 분류표를 이용하였다.

◆ Triphone 모델과 결정트리 기반 상태공유 모델 먼저 본 논문에서 기본적인 모델로 잡고 있는 triphone 모델과 결정트리 기반 상태공유 모델과의 인식 성능 차이를 실험을 통하여 비교한 결과를 표1에 나타내었다.

모델 종류	모델수	상태수	혼합 가우시안수	인식률(%)
Triphone	3,729	11,187	11,187	92.9
Tied state triphone	MG1	3,729	5,019	93.8
	MG2	3,729	5,019	95.6
	MG3	3,729	5,019	96.7
	MG4	3,729	5,019	97.2

표 1 Triphone model과 결정트리 기반 상태공유 모델 간의 인식률 비교

표1에서 MG#은 혼합 가우시안의 수를 나타낸다. 혼합 가우시안의 수를 늘릴수록 인식률은 증가하지만 그 증가율은 적은 반면, 증가하는 파라미터의 수는 그 증가율보다 훨씬 큽을 확인할 수 있었다.

◆ 결정트리 기반 상태공유 모델과 혼합 가우시안 클러스터링 모델

모델 종류	혼합가우시안수	감소율(%)	인식률(%)
Tied state triphone(MG4)	23,898	0	97.2
Euclidean	23,660(238)	1.0	96.94
Bhattacharyya	23,660(238)	1.0	97.2

표 2 결정트리 기반 상태공유 모델과 혼합 가우시안 클러스터링 모델의 성능 비교

감소율이 약 1.0% 지점에서 이전 모델과 거의 동일한 성능을 유지하는 것을 표2를 통해 확인할 수 있다. Euclidean 방법보다는 Bhattacharyya 방법이 조금 더 나은 성능을 보이지만 그 차이는 크지 않음을 알 수 있다.

5. 결론

본 논문은 결정트리 기반 상태공유 모델의 혼합 가우시안 수를 상향식 방법을 이용하여 최적화하고자 하였다. 상향식 방법에 있어서 거리 측정법은 Euclidean과 Bhattacharyya를 이용하였으며 가장 거리가 가까운 두 가우시안으로 새로운 가우시안을 생성하였다. 실험 결과 큰 성능향상을 볼 수는 없었으나 혼합 가우시안 수를 줄여 최적화할 수 있는 가

능성을 확인할 수 있었다.

【참고문헌】

- [1] S.J. Young, J.J. Odell, P.C. Woodland. "Tree-Based State Tying for High Accuracy Acoustic Modelling", ARPA Workshop Human Language Technology, pp. 286-291, Princeton, NJ, Mar. 1994.
- [2] H. J. Nock, M. J. F. Gales, and S. J. Young, "A comparative study of methods for phonetic decision-tree state clustering," Proc. Eurospeech '97, pp. 111-114., Rhodes, Greece, 1997.
- [3] S.J. Young, "The general use of tying in phoneme-based HMM speech recognisers", Acoustics, Speech, and Signal Processing, Vol. 1, pp. 569-572, Mar. 1992.
- [4] Reichl, W., Wu Chou, "Robust decision tree state tying for continuous speech recognition", Speech and Audio Processing, IEEE Transactions on , Vol.8 Issue: 5, pp. 555 -566, Sep. 2000.
- [5] J.J. Odell, "The Use of Context in Large Vocabulary Speech Recognition", PhD's Dissertation. University of Cambridge. 1995.
- [6] Keinosuke Fukunaga, "Introduction to statistical pattern recognition", Morgan Kaufmann, 2nd Edition, 1990.
- [7] Simo O. Kamppari, "Word and Phone Level Acoustic Confidence Scoring for Speech Understanding Systems", B.S., MIT, 1999.
- [8] 오세진, 황철준, 김범국, 정호열, 정현열, "결정트리 상태 클러스터링에 의한 HM-net 구조결정 알고리즘을 이용한 음성인식에 관한 연구", 한국음향학회지, 제21권 제2호, 2002.