1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

음소 모델링 방식들의 성능 비교

송명규, 김형순

부산대학교 전자공학과

Performance Comparison of Acoustic Modeling Techniques

Myung Gyu Song . Hyung Soon Kim

Dept. of Electronics Eng., Pusan National Univ.,

E-mail: {mgsong, kimhs}@hyowon.pusan.ac.kr

요약

HMM 기반의 음성 인식기를 구현하는데 있어서 모델의 복잡도와 제한된 훈련 데이타 사이의 균형을 유지하는 것은 중요한 문제이다. 중간규모 또는 대용량 어휘 인식 시스템은 정교한 모델을 얻기 위해서 문맥종속 음소 모델링이 필수적이다. 그러나, 제한된 훈련 데이터로는 발생 가능한 모든 context 를 포함 하기가 어렵고, 더구나 훈련 데이터에서 관찰된 context 중에서도 그 관찰 빈도가 낮은 것이 많아서 신뢰성 있는 문맥종속 모델들을 얻기에는 여전히 어려움이 따른다. 또한 경우에 따라서는 계산량의 감축을 위하여 모델 규모를 축소시킬 필요도 생긴다. 이러한 문제를 해결하기 위해 본 논문에서는 unit reduction 방법들과 state tying 을 이용한 방법들의 성능을 실험을 통해 비교한다. 고립단어 인식 실험결과 state tying 을 이용한 방법이 unit reduction 에 비하여 우수함을 확인 할 수 있었다.

1. 서론

Hidden Markov Model(HMM)은 시변 음성의 스펙트럼 모델링에 효과적인 것으로 알려져 있다. 그러나, 실제 음성 스펙트럼의 변화를 정확히 모델링 하기 위해서는 많은 모델이 필요하고 또한 비교적 복잡한 출력확률 분포를 사용해야 한다. 이로 인해서 추정해야 할 모델 파라메타가 많아져 훈련 데이터 부족 문제를 야기시킨다.

문맥 종속 모델을 기반으로 하는 시스템은 발생 가능한 triphone 의 수가 많으므로 추정해야 할 모델이 많고, 더구나 훈련 데이터에서 관찰된 triphone 중에서도 그 관찰빈도가 낮은 것이 많다. 이와 같은 이유로 신뢰성 있는 문맥종속 모델을 얻기가 어렵다. 또한 경우에 따라서는 계산량의 감축을 위하여 모델 규모를 축소시킬 필요도 생긴다. 이러한 문제를 해결하기 위한 접근 방법에는 문맥종속 모델의 개수를 줄이는 unit reduction, 각각의 문맥 종속 모델을 모두 가지면서 각 모델의 신뢰도에 따라 적절한 가중치를 부가하는 interpolation, 비슷한 모델 파라메타의 일부를 공유하여 전체 파라메타 수를 줄이는 파라메타 tying 등이 있다.

본 논문에서는 문맥종속 음소 unit 의 관찰 횟수에 따른 unit reduction[1], 통계적 음소 집단화(clustering)[2]에 의한 unit reduction, state tying 을 이용하여 모델 파라메타 수를 줄이는 data-driven clustering[3] 및 decision tree-based clustering[4] 방법의 성능을 실험을 통해 비교한다.

본 논문의 구성은 다음과 같다. 서론에 이어 제 2 장에서는 관찰 횟수 및 통계적 음소 집단화에 의한 unit reduction 방법에 대해 설명하고, 제 3 장에서는 state tying 에 의한 data-driven clustering 및 decision tree-based clustering 방법에 대해서 설명한다. 그리고 제 4 장에서는 실험 방법 및 결과를 기술하고, 제 5 장에서 결론을 맺는다.

2. Unit reduction 방법

Unit reduction 방법은 제한된 음성 데이터를 이용하여 신뢰성 있는 문맥종속 모델을 구성하기 위해서 문맥종속 unit 의 수를 줄여 나가는 방법이다.

2.1 관찰 횟수에 따른 unit reduction

Unit 의 관찰 횟수를 $c(\cdot)$, 신뢰성 있는 모델의 훈련을 위해 필요한 unit 의 관찰 횟수에 대한 threshold 를 T 라

할 때, 사용된 unit reduction 규칙은 다음과 같다[1].

If $c(p_L\text{-}p\text{-}p_R) < T$, then

1. $p_L\text{-}p\text{-}p_R \rightarrow \$\text{-}p\text{-}p_R$ if $c(\$\text{-}p\text{-}p_R) > T$
2. $p_L\text{-}p\text{-}p_R \rightarrow p_L\text{-}p\text{-}\$$ if $c(p_L\text{-}p\text{-}\$) > T$
3. $p_L\text{-}p\text{-}p_R \rightarrow \$\text{-}p\text{-}\$$ otherwise.

여기서 p_L은 left-context Phoneme Like Unit(PLU), p_R은 right-context PLU, $는 don't care condition을 나타낸다. Threshold T를 증가 시키면 전체 모델의 수가 줄고, 문맥독립 음소와 수는 늘어나게 된다.

2.2 통계적 집단화에 의한 unit reduction

이 방법은 각각의 모델들 사이의 거리(distane)를 정의하고, 거리가 작은 모델들 끼리 묶는 방법으로, 단일 mixture를 가지는 모델들 사이의 거리 척도(distance measure)는 다음과 같이 정의하였다.

$$D(p_i, p_j) = \sum_{d=1}^{N} D_d(p_i, p_j) \qquad (1)$$

여기서 p_i, p_j는 각각 i번째와 j번째 모델을 나타내고, N은 모델의 상태 수를 나타낸다. 그리고, $D_d(p_i, p_j)$는 두 모델의 각 상태들 간의 거리로서 다음 식과 같이 주어진다.

$$D_d(p_i, p_j) = \frac{1}{V}\sum_{k=1}^{V}\frac{(\mu_{idk} - \mu_{jdk})^2}{\sigma_{idk}\sigma_{jdk}} \qquad (2)$$

여기서 V는 음성특징벡터의 차원이고, μ_{idk} 및 σ_{idk}는 각각 i번째 음소의 d번째 분포에서 k번째 음성특징 파라메터의 평균 및 표준 편차를 의미한다. 식에 나타난 바와 같이 각각의 모델 내에서의 천이 확률에 의한 영향은 무시하였다.

위와 같은 모델들 사이의 거리 척도(distance measure)를 가지고 유사한 모델들을 집단화 하기 위하여 본 논문에서는 modified k-means(MKM)알고리즘[5]을 사용하였다.

3. State tying을 이용한 방법

앞에서 언급된 unit reduction이 모델 기반의 접근 법으로 좌우 context의 영향을 충분히 반영하지 못하는 반면에, state tying을 이용한 방법은 유사한 상태를 서로 묶어 줌으로서 좌우 context를 보다 잘 반영할 수 있는 방법이다.

3.1 Data-driven clustering

이 방법은 훈련된 triphone 모델의 각 상태를 하나의 독립된 cluster로 두고, 두 cluster 사이의 최소 거리가 threshold보다 클 때까지 최소거리의 두 cluster를 합병하여 전체 상태 수를 줄이는 것이다. 또한, 훈련 데이터가 불충분하여 파라메터 추정의 신뢰성이 떨어질 가능성이 있는 cluster는 가장 가까운 cluster와 합병한다. Data-driven clustering에 사용된 두 상태 사이의 거리는 식 2를 사용한다. 그림 1에 data-driven state tying의 한 예가 나타나 있다.

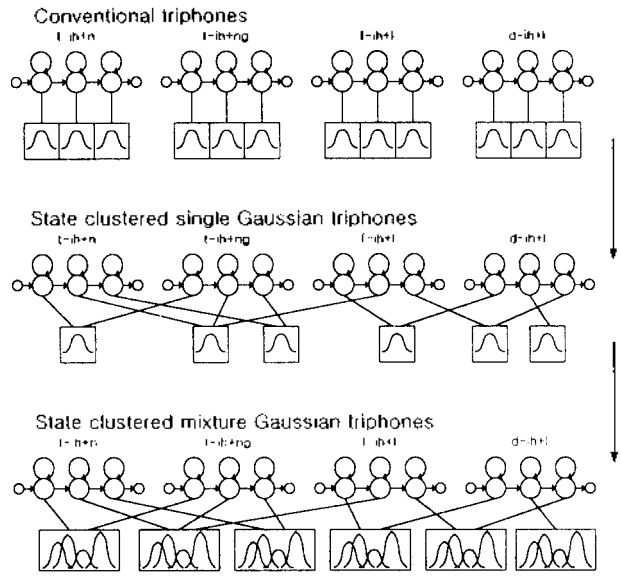

그림 1. data-driven state tying의 예

3.2 Decision tree-based clustering

앞에서 언급된 방법들과는 달리 이 방법은 훈련 데이터에서 관찰되지 않은 문맥종속 모델에 대한 처리가 가능한 방법이다. 이 방법은 음성학적인 decision tree를 이용하여 훈련 데이터에서 관찰되지 않은 문맥종속 모델을 mapping 한다. 음성학적인 decision tree는 각 노드마다 음성학적인 질문이 있는 binary tree이다. Binary tree는 먼저 음성의 기본단위에 해당하는 특정단위(예를 들어 음소 /ㅗ/)에 해당하는 데이터들로 하나의 집합을 구성한 후 이를 두 개의 부분집합으로 나누고 그 각각의 부분집합을 다시 두 개의 부분집합으로 나누어 가

는 일련의 과정을 통해 구성된다. 부분집합으로의 분할은 각 집합에 해당되는 node 에서 음성학적인 binary question 과 binary question 에 대한 평가함수를 필요로 하며 분할을 언제 멈추어야 하는가에 대한 기준도 있어야 한다. Binary question 들의 집합을 Q 라 하고 n 을 tree 에서의 노드, 그리고 $m(q,n)$을 node n 에서 question $q \in Q$ 에 의한 분할에 대한 평가 함수라 할 때 decision tree 구성을 위한 전체적인 알고리즘을 script 언어로 기술하면 다음과 같다.

1. Start with all samples at the root node
2. While there are untested nodes do
   2.1 Select an untested node n
   2.2 Evaluate $m(q,n)$ for all possible questions $q \in Q$ at this node
   2.3 If a stopping criterion is met,
   declare this node as a terminal
   Else
   2.4 associate the question of the highest value of $m(q,n)$ with this node.
   Make two new successor nodes
   All samples that answer positively to the question are transferred to the left successor and the others to the right successor

그림 2 는 3 개의 상태를 가지는 음소 /ㅗ/에 대한 모델들의 가운데 상태를 decision tree-based clustering 에 의해 5 개의 cluster 로 나타내는 예이다.

Decision tree-based clustering 은 훈련 데이터에서 관찰되지 않은 문맥종속 모델에 대한 문제를 해결하면서도 data-driven clustering 과 비슷한 성능을 나타내는 것으로 보고되었다[4].

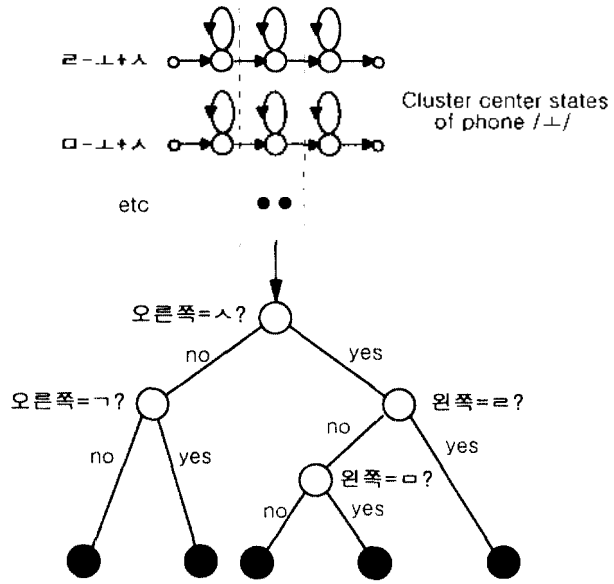

그림 2. Decision tree-based state tying 과정

4. 실험 및 결과

4.1 데이터 베이스 및 인식 시스템

인식 실험을 위해 음성 다이얼링을 목적으로 세 가지의 다른 환경에서 수집된 한국통신의 전화음성 데이터베이스를 이용하였다. 각각의 데이터 베이스는 VDS, VDS95, VDSfromREAL1 로 명명되어 있다 음성 신호는 8kHz 로 샘플링 되었으며 8bit μ-*law* PCM 으로 부호화되어 있다.

인식 시스템은 150 개의 고립 단어를 인식하도록 구성하였다. 연속 확률분포를 갖는 HMM 의 훈련을 위해서는 VDS 와 VDS95 DB 의 일부를 사용하였고, 인식 실험을 위해서는 VDS, VDS95 DB 중 훈련에 참가하지 않은 나머지 데이터와, VDSfromREAL1 DB 를 사용하였다.

입력음성은 $1-0.97z^{-1}$의 전달함수를 갖는 필터로 preemphasis 되고, 이 음성은 프레임 단위로 분할되어 처리 되는데 각 프레임은 20msec 의 길이를 가지는 hamming 윈도우를 10msec 마다 씌워서 얻어진다. 분할된 각 프레임에 26 차의 필터뱅크를 이용하는 MFCC 분석과정을 거쳐 멜-켑스트럼 계수를 얻는다. 음성특징 벡터는 12 차의 가중 멜-켑스트럼과 그것들의 일차빼기, 이차빼기, 로그파워 그리고 로그파워의 일차 및 이차빼기의 총 39 차로 구성된다. 음소 모델은 상태 당 하나의 mixture 를 갖는 3 상태로 모델링 하였으며, 묵음 모델은 상태 당 5 개의 mixture 를 갖도록 하였다.

4.2 인식 실험 및 결과

제한된 음성 데이터를 이용하여 신뢰성 있는 문맥종

속 음소모델을 구성하기 위한 unit reduction 및 파라메터 tying의 실험결과가 표 1에 나타나 있다. 이 실험은 훈련 데이터에서 관찰된 546 개의 biphone 및 triphone 모델을 약 150 개(450 개 상태)로 줄인 경우의 실험 결과이다.

표 1. Baseline, unit reduction 및 state tying 방법의 인식률

	상태 수	VDS	VDS95	VDSfR1	total
Baseline	1638	94.9%	91.4%	91.9%	92.7%
Unit reduction by count	450	79.1%	77.4%	69.4%	75.4%
Unit reduction by clustering	450	84.8%	86.8%	77.5%	83.6%
Data-driven clustering	450	92.4%	91.2%	90.4%	91.4%
Tree-based clustering	450	90.4%	85.7%	89.3%	88.3%

위 표에서 Baseline은 훈련 데이터에서 관찰된 546 개의 biphone 및 triphone 모델을 모두 이용하여 인식실험을 한 결과이다. 참고로 본 논문에서의 모든 실험에는 전화망을 통한 채널왜곡에 대한 처리를 도입하지 않았으며, 이를 적용한 경우 추가적인 성능향상이 기대된다. 관찰 횟수에 의한 unit reduction은 상당히 저조한 성능을 보였고, 통계적 집단화에 의한 unit reduction은 관찰 횟수에 의한 unit reduction 방법보다는 나은 성능을 보이지만, 여전히 state tying 방법들에 비하여 저조한 성능을 보였다. 이미 언급한 바와 같이 state tying 방법은 좌우의 context를 고려할 수 있는 반면에 unit reduction 방법은 모델단위로 묶어 줌으로써 좌우 context를 제대로 반영하지 못하여 저조한 성능을 나타낸 것으로 판단된다. State tying 방법 중에서는 data-driven clustering 방법이 decision tree-based clustering 방법보다 나은 성능을 보였다. Data-driven clustering 방법은 전체 파라메터 수를 70%이상 줄였는데도 전체 인식률은 1.3%만 저하되는 성능을 보였다. 그러나 data-driven clustering 방법은 decision tree-based clustering 방법과는 달리 훈련 데이터에서 관찰되지 않은 문맥종속 모델에 대한 대치가 불가능 하다는 문제점이 있다.

5. 결론

본 논문에서는 제한된 음성데이타를 이용하여 계산량과 정확도가 고려된 효과적인 음소모델링 방법을 살펴보았다. 모델 tying에 비해서 파라메터 tying이 우수한 성능을 나타냄을 확인할 수 있었고, 주어진 훈련 데이터에 가장 적합한 문맥종속 모델은 data-driven clustering에 의해 얻을 수 있었다. 그러나, 훈련 데이터에서 관측되지 않은 context를 처리하려면 decision tree-based clustering 방법을 사용할 필요가 있다.

본 연구 결과를 기반으로 하여 화자 잡음제거 및 거절기능 구현을 위한 연구가 계속 진행중이다.

참 고 문 헌

[1] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[2] 이활림, 김형순 외, "음소 HMM 모델을 이용한 keyword spotting 시스템에서의 non-keyword 모델에 관한 연구," 제 12 회 음성통신 및 신호처리 워크샵 논문집, pp.83-87, 1995 년.

[3] S. J. Young, et al, *HTK: Hidden Markov Model Toolkit V2.0*, Entropic Cambridge Research Laboratory, 1995

[4] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modelling," in Proc. ARPA Human Language Technology Workshop, pp.286-291, 1994

[5] J. G. Wilpon and L. R. Rabiner, "A Modified K-means Clustering Algorithm for Use in Isolated Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol.33, no.3, pp.587-594, June 1985.

본 연구는 1998 년 한국통신 장기기초 연구과제의 연구비 지원을 통하여 이루어 졌습니다.