High-Performance 음성 인식을 위한 Efficient Mixture Gaussian 합성에 관한 연구

이상복, 이철희, 김종교
전북대학교 전자정보공학부
전화 : (063) 272-1177 / 팩스 : (063) 270-2400

A Study on Gaussian Mixture Synthesis for High-Performance Speech Recognition

Sang-Bok Lee, Chul-Hee Lee, Chong-Kyo Kim
Dept. of Electronics and Information Eng., Chonbuk National University

Abstract

We propose an efficient mixture Gaussian synthesis method for decision tree based state tying that produces better context-dependent models in a short period of training time. This method makes it possible to handle mixture Gaussian HMMs in decision tree based state tying algorithm, and provides higher recognition performance compared to the conventional HMM training procedure using decision tree based state tying on single Gaussian GMMs. This method also reduces the steps of HMM training procedure. We applied this method to training of PBS, and we expect to achieve a little point improvement in phoneme accuarcy and reduction in training time.

1. 서 론

ASR 시스템은 보다 높은 인식 성능을 위해 mixture Gaussian CD(Context Dependent)모델을 사용한다[1]. 일반적으로 구성된 triphone HMM(Hidden Markov Model)에서 decision tree based state tying을 수행하였다[2]. State tying 과정을 거쳐 모아진 HMM state들은 single Gaussian HMM 방식과 mixture Gaussian HMM 방식이 있는데, single Gaussian HMM 방식에서는 음소 단위의 음향학적인 특징을 표현하기에 미흡하다는 점과 training 과정에서 상당한 시간을 요구하는 문제가 따르게 된다.

따라서, 본 논문에서는 decision tree based state tying에서 mixture Gaussian HMM을 다루는 효율적인 방법을 제안한다. 제안한 방법은 HMM state들에서 mixture 개수만큼의 Gaussian 들이 나오고 이를 N 개로 clustering을 해서 다시 합성 과정을 거치게 된다.

본 논문의 구성은 2절에서 state tying에 대해 알아보고 3절에서는 state tying 과정에서 mixture Gaussian 합성법에 관해 설명하고 4절에서 실험을 하여 기존 방법과 비교 한 후, 5절에서 결론을 맺는다.

2. Decision Tree based State Tying

음소 모델은 훈련 데이터가 충분한 경우, 그 음소의 좌우 문맥을 고려한 문맥 종속형(Context-Dependent) 모델을 사용하면 음소간의 조음화(coarticulation)현상을 보다 효과적으로 반영할 수 있기 때문에 인식률의 향상을 가져온다. 여기에는 음소의 개수에 따라 bi-phone, triphone, quinphone 등이 있는데, 본 논문에서는 word-internal triphone 을 사용하였다. 구성된 triphone 모델의 각 상태들은 decision tree의 상위 root node로 모아진다. 그리고 아래로 진행되면서 node는 두 개의 node로 분리된다. 이때 각 노드마다 음운학적

인 질문을 통해 최종 leaf 노드로 clustering된다[3]. 그림 1에 clustering에 쓰일 decision tree를 나타냈다. 이 과정은 분리된 두 노드의 log likelihood의 합과 분리 전 노드의 log likelihood와의 차가 일정 threshold보다 작을 때까지 진행된다. 이러한 state tying을 통해서 훈련과정에서 거의 나타나지 않는 unseen 모델을 해결할 수 있게 된다[4].

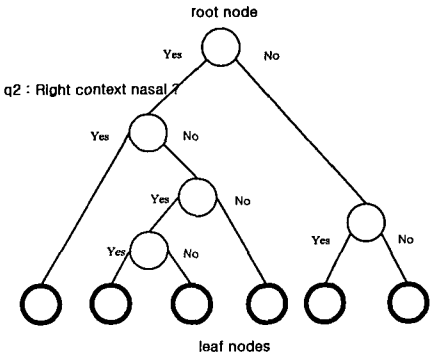

그림 3. 음소 decision tree

Node S_m이 질문 q에 의해 $S_{m,y}(q)$, $S_{m,n}(q)$ 로 나누어진다고 가정하면, log likelihood의 변화값은 식(1)과 같다.

$$\Delta L_q = L(S_{m,y}(q)) + L(S_{m,n}(q)) - L(S_m) \quad (1)$$

그리고 node S_m의 mean vector 와 covariance matrix의 계산은 식(2)와 식(3)을 통해 구할 수 있다.

$$\mu_m^{(k)} = \sum_i \Gamma_{m,i}\mu_{m,i}^{(k)} \Big/ \sum_i \Gamma_{m,i} \quad (2)$$

$$\sigma_m^{(k)} = \Big[\sum_i \Gamma_{m,i}(\mu_{m,i}^{(k)} - \mu_m^{(k)})^2 + \sum_i \Gamma_{m,i}\sigma_{m,i}^{(k)}\Big] \Big/ \sum_i \Gamma_{m,i} \quad (3)$$

여기서, $\mu_{m,i}$, $\sigma_{m,i}$, $\Gamma_{m,i}$ 는 각각 평균벡터, 공분산 행렬, node S_m 의 i 번째 상태가 차지하는 프레임수의 기대치를 가리킨다. 숫자 k 는 특징 벡터의 k 번째 성분을 나타낸다. 훈련 벡터의 관측열 $O_t(t=1,2,\cdots,T)$ 이 주어지면, node S_m 의 log likelihood 값은 다음으로 주어진다.

$$L(S_m) \approx \sum_{t=1}^{T} \log[N(O_t, \mu_m, \sigma_m)] \cdot \gamma_t(m) = -\frac{1}{2}(K\log 2\pi + \log|\mu_m| + K)\Gamma_m \quad (4)$$

여기서, $\gamma_t(m)$, Γ_m은 각각 t 시간에 node S_m 에 있는 상태들이 점유될 확률, 관측열에서 차지하는 프레임 수의 기대값을 가리킨다. 그러므로 $N(O_t, \mu_m, \sigma_m)$은 평균 벡터 μ_m 과 공분산 행렬 σ_m 의 Gaussian 분포가 관측열 O_t 를 출력할 확률을 나타낸다.

Decision tree based state tying은 결과가 single Gaussian HMM이므로 mixture Gaussian HMM을 얻으려면 이후에 mixture 증가 과정과 파라미터 재추정 과정을 추가적으로 반복해야 한다. Mixture 수의 증가는 두 개로 분할하고 mixture weight 값을 곱해줌으로써 얻게 된다. 이러한 과정을 원하는 성능이 나오거나 원하던 mixture 수를 얻기까지 반복해야 한다.

3. Mixture Gaussian 의 합성

식(4)에서는 state tying을 결정짓는 log likelihood $L(S_m)$이 single Gaussian 분포에서 계산된다. 하지만, mixture Gaussian 분포는 single Gaussian 분포에 비해 음향학적 특성을 더 정확히 반영함으로써, 좋은 인식 성능을 나타낸다. 그러므로 이번 실험은 mixture Gaussian 기반에서 state tying 과정을 하게 된다.

본 논문에서 제안한 방법은 tying 과정을 거치지 않은 triphone이나 tying 과정을 거친 triphone에 대해서도 mixture Gaussian 분포를 다루고, clustering과정에서 분포들이 Gaussian 분포임을 가정한다. Decision tree에서 각 노드에 있는 상태들을 구성하는 Gaussian 분포가 N 개로 모아진다. 여기서, N 은 노드 안의 state들 중에서 가장 많은 개수의 mixture 수가 된다. 모아진 mixture Gaussian 분포는 다시 하나의 Gaussian으로 합성된다. 기존 방법과 제안한 방법을 비교하면, 그림 2와 같다.

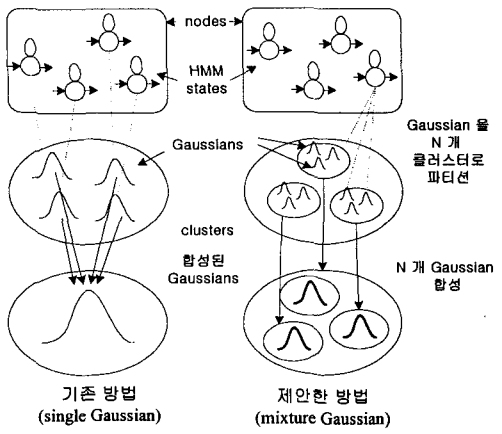

그림 4. 기존 방법과 제안한 방법과의 비교

N 개의 class들로 모아지는 과정에서 K-means clustering 알고리즘을 사용한다. 입력 모델의 평균값은 그 요소들이 되고, clustering의 distance metric은 분산값이 곱해진 Euclidean distance에 해당한다. 각 class 별로 Gaussian으로 합성된다. 그 중 N 번째에 해당하는 합성된 Gaussian의 평균값과 공분산값, 그리고 합성된 Gaussian의 mixture weight 값은 아래 식과 같이 주어진다.

$$\mu_{m,n}^{(k)} = \sum_i \Gamma_{m,n,i}\, \mu_{m,n,i}^{(k)} \;/\; \sum_i \Gamma_{m,n,i} \tag{5}$$

$$\sigma_{m,n}^{(k)} = \Big[\sum \Gamma_{m,n,i} (\mu_{m,n,i}^{(k)} - \mu_{m,n}^{(k)})^2 + \sum_i \Gamma_{m,n,i} \sigma_{m,n,i}^{(k)} \Big] / \sum_i \Gamma_{m,n,i} \tag{6}$$

$$\omega_{m,n} = \sum_i \Gamma_{m,n,i} / \sum_n \sum_i \Gamma_{m,n,i} \tag{7}$$

n 번째 class에서 i 번째 성분이 차지하는 프레임 수의 기대값인 $\Gamma_{m,n,i}$ 는 원래 state가 차지하는 기대 프레임수와 mixture weight 값과의 곱으로 근사화된다. 이렇게 합성된 N 개의 합성 Gaussian은 훈련 데이터의 log likelihood를 계산하는데 쓰인다. Gaussian 분포의 overlap을 사용하여 log likelihood값을 계산하는데, 이때 많은 시간이 요구된다. 따라서 본 논문에서는 overlap을 무시하여 log likelihod값을 근사화하는 방법을 제안함으로써, log likelihood 값을 근사화한다 [5]. 식(4) 대신 node S_m 의 log likelihood값은 다음과 같이 근사화된다.

$$\begin{aligned} L(S_m) &\approx \sum_{t=1}^{N} \log\Big[\sum_{n=1}^{N} \omega_{m,n} \mathcal{N}(O_t, \mu_{m,n}, \sigma_{m,n}) \Big] \cdot \gamma_t(m) \\ &\approx \sum_{t=1}^{T} \log[\max\{\omega_{m,n} \mathcal{N}(O_t, \mu_{m,n}, \sigma_{m,n})\}] \cdot \gamma_t(m) \\ &\approx \sum_{n=1}^{N} \Big[\Gamma_{m,n} \log \omega_{m,n} - \frac{\Gamma_{m,n}}{2} (K \log 2\pi + \log|\sigma_{m,n}| + K) \Big] \\ &\approx \sum_{n=1}^{N} \Big[\Gamma_{m,n} \log \Gamma_{m,n} - \frac{\Gamma_{m,n}}{2} (K \log 2\pi + K + \log|\sigma_{m,n}|) \Big] - \sum_{n=1}^{N} \Gamma_{m,n} \cdot \log \sum_{n=1}^{N} \Gamma_{m,n} \end{aligned} \tag{8}$$

여기서, $\mu_{m,n}$, $\sigma_{m,n}$, $\omega_{m,n}$, $\Gamma_{m,n}$ 은 각각 평균벡터, 공분산 벡터, mixture weight, node S_m 의 n 번째 합성된 Gaussian 분포가 차지하는 기대 프레임 수를 나타낸다. 제안된 state tying 과정의 전체 흐름은 그림 3 과 같다.

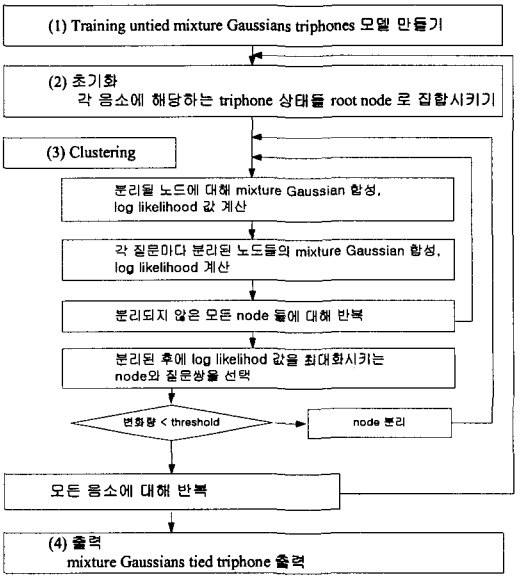

그림 5. State tying 과정

4. 실험 결과

제안한 방법과 기존의 방법으로 훈련되어진 triphone 모델의 성능은 PBS(Phonetically Balanced Sentence)를 이용하여 비교하였다. 기존 방법의 모델 훈련 과정은 다음과 같은 과정으로 진행된다.

1. Single Gaussian 의 untied triphone 구성
2. 기존의 결정트리 state tying 수행
3. 파라미터 재추정 과정 3번 반복
4. Mixture 수 2개로 증가
5. 파라미터 재추정 과정 3번 반복
6. Mixture 수 4개로 증가
7. 파라미터 재추정 과정 3번 반복
8. Mixture 수 8개로 증가
9. 파라미터 재추정 과정 3번 반복

2, 4, 8개 mixture의 triphone 모델은 위의 과정으로 얻어진다. 본 논문에서 제안한 방법은 다음과 같다.

1. Mixture Gaussians 의 untied triphone 구성
2. 제안한 state tying 수행
3. 파라미터 재추정 과정 3회 반복

2, 4, 8개 mixture의 triphone 모델은 개별적인 과정으로 얻어진다. 실험에 대한 평가는 decision tree state tying 직후의 triphone 모델과 3번의 파라미터 재추정 과정 직후의 triphone 모델에서 이루어졌다. 훈련용 데이터로는 7명의 남성 화자와 7명의 여성 화자의 1,400 문장을 사용하였고, 특징 파라미터로는 12차 MFCC 및 delta 계수, acceleration 계수를 사용하였다. 남성 화자와 여성 화자는 각각 독립적으로 평가하였는데, 평가용 데이터로는 3명의 남성 화자와 3명의 여성 화자의 총 600문장을 사용하였다. 표 1과 표 2에 기존 방법과 제안한 방법의 결과를 남성 모델과 여성 모델로 구분해서 인식률을 비교하였다. Baseline 구성에서 mixture splitting을 통해 4개까지 모델을 구성한 결과와 제안한 방법을 통해 얻은 모델과의 인식률 결과를 보면, 평균적으로 1~2 point 인식률의 향상을 확인할 수 있었다. 또한, 훈련 스텝 수의 감소로 훈련 시간을 단축함을 알 수 있었다.

표 1. 남성 모델

	iteration	기존 방법 인식률(%)	제안한 방법 인식률(%)
baseline	0	66.5	·
	3	66.6	·
mixture 2	0	67.2	68.6
	3	68.5	69.5
mixture 4	0	69.4	71.8
	3	70.9	72.2

표 2. 여성 모델

	iteration	기존 방법 인식률(%)	제안한 방법 인식률(%)
baseline	0	64.7	·
	3	65.3	·
mixture 2	0	65.7	67.7
	3	67.1	68.4
mixture 4	0	67.6	69.4
	3	69.3	70.2

5. 결 론

본 논문은 decision tree state tying 과정에서 mixture Gaussian 기반의 효율적인 방법을 제시하였다. 이 방법은 우리가 목표로 하는 mixture 수의 모델에 대해 보다 뛰어난 state tying을 제공하고 mixture Gaussian HMM 을 다룰 수 있게 한다. 제안한 방법은 기존 방법에 비해 1~2 point의 인식률 향상과 훈련 시간의 감소를 이룰 수 있었다.

참고 문헌

[1] Lawrence Rabiner, Biing-Hwang Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[2] W. Reichl and W. Chou, "Decision tree state tying based on segmental clustering for acoustic modeling," *Proc. ICASSP 98*, pp. 801-804, 1998.

[3] D. Willett et al., "Refining tree-based state clustering by means of formal concept analysis, balanced decision trees and automatically generated model-sets," *Proc. ICASSP 99*, pp. 565-568, 1999.

[4] L. R. Bahl et al., "Decision trees for phonological rules in continuous speech," *Proc ICASSP 91*, pp. 185-188, 1991.

[5] Tsuneo Kato et al., "Efficient mixture Gaussian synthesis for decision tree based state tying," *Proc. ICASSP 2001*, pp. 493-496, 2001.

[6] Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho, Phil Woodland, *The HTK Book* (for HTK Version 2.2), Entropic Cambridge Research Laboratory, 1999.