논문 2005-42SP-6-19

혼합 가우시안 군집화를 이용한 상태공유 음향모델 최적화

(A Study on the Optimization of State Tying Acoustic Models using Mixture Gaussian Clustering)

안 태 옥*

(Tae Ock Ann)

요 약

본 논문은 음성인식에 쓰이는 음향모델의 모델링 방법 중 결정트리 상태공유 모델링(DTST)을 기반으로 출력 확률 분포의 혼합 가우시안 수를 줄여 모델을 최적화하는 방법을 제안한다. DTST는 음성학적 지식을 포함할 수 있는 질의어 집합과 유사도를 기반으로 한 결정 방법을 이용하는 것이다. 이때 상태들의 출력 확률 분포의 혼합 가우시안 수를 늘려 인식률을 증가시킬 수 있게 된다. 본 논문에서는 인식률이 최대가 되는 지점에서 혼합 가우시안들을 군집화 하여 그 수를 줄이고자 한다. 군집화 시에 필요한 거리 측정 방법은 유클리드(Euclidean)와 바타챠랴(Bhattacharyya) 방법을 이용하였고, 새로운 가우시안은 거리가 최소가 되는 두 가우시안으로부터 평균과 분산을 다시 계산하여 생성하였다. 증권상장 회사명(STOCKNAME) 1,680개의 단어 데이터베이스를 구성하여 실험한 결과 바타챠랴 방법은 97.2 %의 인식률을 유지하면서 전체 혼합 가우시안 수의 비율을 1.0 %로 감소시켰고, 유클리드 방법은 96.9 %의 인식률을 유지하면서 혼합 가우시안 수의 비율을 1.0 %로 감소시켜 모델을 최적화할 수 있었다.

Abstract

This paper describes how the state tying model based on the decision tree which is one of Acoustic models used for speech recognition optimizes the model by reducing the number of mixture Gaussians of the output probability distribution. The state tying modeling uses a finite set of questions which is possible to include the phonological knowledge and the likelihood based decision criteria. And the recognition rate can be improved by increasing the number of mixture Gaussians of the output probability distribution. In this paper, we'll reduce the number of mixture Gaussians at the highest point of recognition rate by clustering the Gaussians. Bhattacharyya and Euclidean method will be used for the distance measure needed when clustering. And after calculating the mean and variance between the pair of lowest distance, the new Gaussians are created. The parameters for the new Gaussians are derived from the parameters of the Gaussians from which it is born. Experiments have been performed using the STOCKNAME (1,680) databases. And the test results show that the proposed method using Bhattacharyya distance measure maintains their recognition rate at 97.2 % and reduces the ratio of the number of mixture Gaussians by 1.0 %. And the method using Euclidean distance measure shows that it maintains the recognition rate at 96.9 % and reduces the ratio of the number of mixture Gaussians by 1.0 %. Then the methods can optimize the state tying model.

Keywords: Speech Recognition, Signal Processing, Acoustic Model , State Tying, Clustering,

I. 서 론

음성인식은 모델과의 비교를 통해 인식 대상을 판단하는 것이다. 음성 인식에서는 인식 대상 어휘를 대상으로 모델을 생성하며 모델의 형태는 은닉 마코프 모델(Hidden Markov Models)[3]이다. 이 모델은 확률적으로 처리하는 기법으로 샘플들을 기반으로 필요한 확률 파라미터를 추정하게 된다. 이러한 샘플들은 특징 파라미터로써 신호 처리 기법을 이용하여 추출하며 순수하게 이 특징 파라미터들로만 이루어진 모델을 음향 모델이

* 정회원, 호원대학교 컴퓨터학부
(Division of Computer, Howon Univ.)
※ 본 논문은 호원대학교 교내 연구비에 의한 연구 논문임
접수일자:2005년7월27일, 수정완료일: 2005년9월26일

라 한다[2]. 음향 모델링 방법은 모델의 강건함과 음성의 변화를 고려한 모델의 다양화 사이의 균형이 필요하다[1]. 두 가지를 모두 반영하기 위해 여러 공유 모델링 기법들이 연구되어 왔다. 공유 형태를 지니는 모델링 기법은 크게 상향식(bottom up) 방법과 하향식(top down) 방법이 있다. 상향식 방법은 모델의 상태 수가 하나인 것에서부터 시작하여 다수의 상태들을 분할하는 방법[13]이고. 하향식 방법은 트라이폰 (triphone) 모델을 음향적인 성질이 비슷한 것들로 묶은 (generalized triphone) 기법[5], 모델보다 하위 수준의 상태를 음성학적 유사성과 최대 확률 기법을 이용한 상태공유(state tying) 기법[6], 상태의 출력 확률 분포가 비슷한 것들을 공유하는 분포공유(distribution tying) 기법[7], 그리고 분포를 결정짓는 파라미터를 공유하는 특징공유(feature tying) 기법[4]등이 있다. 이러한 연구방법들은 시간이 지남에 따라 공유 수준이 점점 세분화되고 있다.

본 논문의 공유 기법은 결정트리를 이용한 상태공유 기법을 이용한다. 이 방법은 원하는 언어의 음향적인 특성과 음성학적인 지식을 최대 확률 기법에 따라 모델에 효과적으로 반영할 수 있다. 이 기법이 갖는 장점은 첫째, 결정트리의 분류와 예측으로 훈련데이터에서 나타나지 않은 모델의 합성을 가능하게 한다. 둘째, 결정트리 기반의 상태공유의 노드 분할 과정과 모델 선택 과정을 통해 모델의 복잡성을 완화하고, 제한된 훈련 데이터로부터 강건한 모델 파라미터 추정이 가능하게 되어 필요한 파라미터 양과의 균형을 유지할 수 있다[8]. 이러한 장점을 가진 기법으로 모델을 생성한 후, 각 상태들의 관측 분포(가우시안 분포)를 다시 분할하여 좀 더 세밀한 모델링을 하는 것을 혼합 가우시안(mixture Gaussian)을 생성한다고 한다[1]. 만약 혼합 가우시안의 수를 M개로 한다고 하면, 각 관측 분포의 파라미터의 수는 상태 수의 M배로 증가하게 된다. 관측 분포의 세분화에 따라 인식률 역시 향상하게 된다. 하지만 그 수가 정비례하여 늘어나게 되므로 본 논문에서는 생성된 혼합 가우시안을 군집화 하여 혼합 가우시안의 수를 줄임으로써 상태공유 모델을 좀 더 최적화 하고자 한다. 가우시안을 군집화 하기 위해서는 거리 측정 기법과 합성 기법이 필요하게 되는데 본 논문에서는 Euclidean 과 Bhattacharyya 거리 측정법 을 비교하기 위해 사용하였다[10][11].

본 논문의 구성은 I 장 서론에 이어, II장에서는 음향모델링기법에 대해 알아보고, III장에서는 음향모델의 파라미터공유 에 대해 알아보고, IV장에서는 제안한 혼합 가우시안 군집화를 이용한 모델의 최적화 방법에 대해 알아보고, V 장에서는 실험결과를 통한 제안한방법의 유효성을 확인하며, VI장에서 결론을 맺는다.

II. 음향모델링 기법

1.음향모델링 단위 정의

음소는 문맥에 따른 음향적 변화가 매우크기 때문에 음향모델단위로 적절하지 못 하다.이에 반해 단어 모델은 음향적 표현이 아주 잘 정의 되고, 단어 사전(lexicon)을 만들 필요가 없기 때문에 인식 구조를 간단하게 해주는 이점을 지니고 있다. 그러나 신뢰할 만한 단어 모델을 얻기 위해서는 훈련 집합에서 단어 발화의 수가 충분히 많아야하고, 각 단어들의 음성학적 내용은 대 어휘에서 반드시 겹치게 되어 단어 모델의 학습 과 인식수행과정에서 과도한 부하를 초래하게 된다. 이러한 이유로 좀 더 효율적인 음성표현으로 단위음소 를

표 1. 기본 PLUs 집합
Table 1. Set of basic phonelike units (PLUs).

Number	Symbol	Position	Phone	Number	Symbol	Position	Phone
1	sil		silence	23	TT	초성	ㄸ[t͈]
2	K	초성	ㄱ[g]	24	PP	초성	ㅃ[p͈]
3	KQ	종성		25	SS	초성	ㅆ[s͈]
4	N	초성	ㄴ[n]	26	ZZ	초성	ㅉ[tɕ͈]
5	NQ	종성		27	AA	-	ㅏ[ɑ]
6	T	초성	ㄷ[d]	28	AX	-	ㅓ[ʌ]
7	TQ	종성		29	OW	-	ㅗ[o]
8	L	초성	ㄹ[l]	30	UW	-	ㅜ[u]
9	LQ	종성		31	WW	-	ㅡ[ɯ]
10	M	초성	ㅁ[m]	32	IY	-	ㅣ[i]
11	MQ	종성		33	E	-	ㅔ[e]
12	P	초성	ㅂ[b]	34	JA	-	ㅑ[ya]
13	PQ	종성		35	JX	-	ㅕ[y]
14	S	초성	ㅅ[s]	36	JO	-	ㅛ[yo]
15	NX	종성	ㅇ[ŋ]	37	JU	-	ㅠ[yu]
16	Z	초성	ㅈ[dʑ]	38	JE	-	ㅖ[ye]
17	CH	초성	ㅊ[tɕ͈]	39	WA	-	ㅘ[wa]
18	KH	초성	ㅋ[kʰ]	40	WX	-	ㅝ[wə]
19	TH	초성	ㅌ[tʰ]	41	UI	-	ㅟ[ɥ]
20	PH	초성	ㅍ[pʰ]	42	WE	-	ㅙ[ɵ]
21	HH	초성	ㅎ[h]	43	WI	-	ㅢ[ɨy]
22	KK	초성	ㄲ[k͈]				

사용하게 된다. 가장 대표적인 단위는 PLUs(Phonelike units)이다[2]. 이는 소리의 기본적인 음소 집합을 사용하여 음향적 유사성을 기반으로 단위를 정하게 된다. 표 1은 본 논문에서 사용한 PLUs를 나타내고 있다.

PLUs는 문맥 독립적인 모델이라고 할 수 있다. 이러한 모델의 이점은 문맥 독립적인 하위 단어(subword)의 작은 기본 집합을 사용함 으로써 음성데이터베이스로부터 쉽게 훈련될 수 있고 , 별도의 수고 없이 새로운 문맥으로 일반화 시킬 수 있으며, 모델을 추출된 훈련 토큰들(training tokens)로부터 문맥의 세밀함에 상대적으로 영향을 덜 받게 할 수 있다

그러나 문맥 독립적인 PLUs 훈련 집합은 문맥종속적인 훈련집합 보다 낮은 인식 성능을 보인다. 이러한 성능 차이는 문맥 독립적인 하위 단어 단위 모델들이 모든 문맥 내의 음성 단위의 스펙트럼과 일시적인 특성을 표현하는데 충분하지 않기 때문이다. 이러한 문제를 해결하는 방법은 인식 시스템에 문맥 종속적인 단위를 포함시키도록 하여 하위 단어 단위의 집합을 확장하는 것이다. 다음은 PLUs의 기본 집합을 트라이폰 으로 확장한 형태를 보여주고 있다.

$p_L - p - p_R$ (Left-right context: LRC) :triphone

p: PLUs에서 정의된 임의의 음소를 나타낸다.

Ⅲ. 파라미터 공유

트라이폰 모델을 기본 단위로 사용할 경우 표1의 PLUs를 기반으로 할 때 42 * 42 * 42 = 74,088개의 모델이 필요하게 된다. 모델의 수도 많지만 모델을 강건하게 만들기 위해서는 음성데이터의 양 역시 많이 필요하게 된다. 이러한 문제에서 음향 모델링의 연구에서 공유 구조(tied structure)는 중요한 역할을 한다. 이는 generalized triphone[5],state tying[6], distribution tying[7], feature tying[4] 에서 보여 진다. 공유 구조를 이용하여 생성하는 주요 문제는 최소의 복잡성으로 데이터의 특성을 표현하는 일반적인 모델을 어떻게 얻느냐하는 데에 있다. 공유 구조를 사용함으로써 증가하는 파라미터를 줄일 수 있어 훈련의 효율성을 증대시킬 수 있고, 모델 파라미터의 수를 줄임으로써 인식에 필요한 계산 양을 줄일 수 있다. 공유 수준에 따라 모델(model), 상태(state), 분포(distribution), 특징(feature) 수준으로 나눌 수 있다.

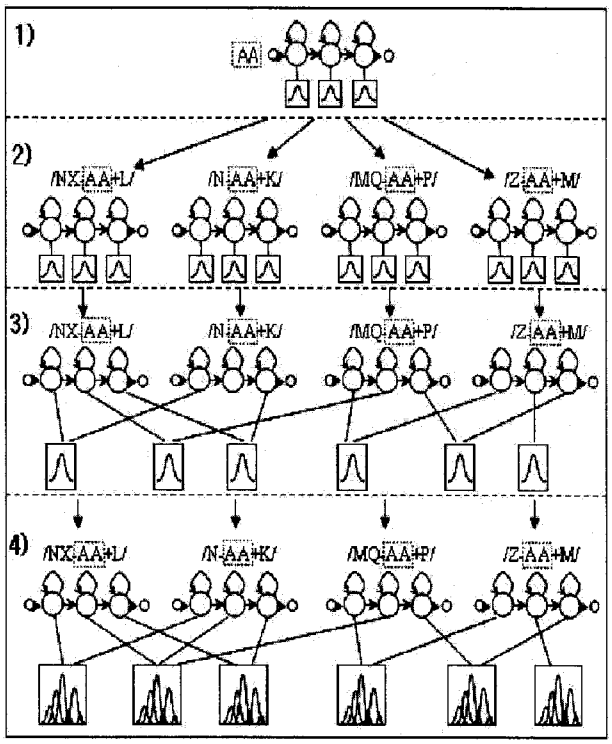

그림 1. 상태공유 HMM 생성 과정
Fig. 1. The Tied-state HMM build procedure.

1. 결정트리 기반 상태공유 모델

결정트리 기반 상태공유 음향 모델링의 목적은 언어의 음향적, 음성학적인 지식을 일정한 최대 확률 기법에 따라 모델에 효과적으로 포함시키는데 있다. 그림1은 상태공유 HMM모델 을 생성하는 과정을 나타내고 있다.수행과정은 다음과 같이 네 단계로 나뉘어 수행된다. 1) 단일 가우시안 출력 확률 밀도 함수를 갖는 3상태 단 음소(3 state left-right monophone) 모델 초기집합을 생성하고 훈련한다. 2) 단 음소의 상태 출력 분포는 Baum-Welch 재 추정을 사용하여 훈련된 트라이폰 모델집합을 초기화하기 위해 복사한다.(전이확률 행렬(transition probability matrix)은 복사되지 않지만 각 음소의 모든 트라이폰이 공유하게 한다) 3) 동일한 단 음소로부터 유도된 트라이폰들의 각 집합에 대해 대응되는 상태들을 군집화 한다. 각 결과 군집에서 대표적인 상태가 선택되고 모든 군집 내의 상태들은 대표 상태로 묶이게 된다. 4) 각 상태의 혼합 요소의 수를 증가시켜 재추정하여 모델의 정밀도를 향상 시킨다

2. 결정트리기반 음향모델 군집화

음성학적 결정트리는 각 노드에 질의(question)가 첨부된 이진(binary) 트리이다. 질의어들은 중심상태의 왼쪽이나 오른쪽으로 음성학적 문맥과 연관되어 있다. 이런 트리는 훈련 데이터에 없던 트라이폰 모델을 합성해 낼 수 있고, 해당 트라이폰 모델의 문맥과 가장 비슷한 단말 트리 노드를 찾아 이와 연관된 공유상태들로 구성할 수 있게 한다. 모든 질의는 "왼쪽 혹은 오른쪽 음소가 집합 X의 원소인가?"의 형태이다. 이때 집합 X는 비음(nasal), 마찰음(fricative), 모음(vowel) 등과 같은 광범위한 것에서 단일 집합 {l}, {m} 등과 같이 단순한 음성학적 분류로 이루어진다. 각 트리는 하향식으로 순차적 최적화 과정을 통해 생성된다[9]. 처음에 군집화 된 모든 상태들이 트리의 루트 노드로 대치되며, 그 노드 내에서 모든 상태들이 공유됐다는 가정 하에 훈련 데이터의 로그 확률을 계산한다. 이 노드는 로그 확률을 최대로 증가시키는 부모 노드에서 상태들을 나누는 질의를 찾음으로써 두 개로 분할하게 된다. (이 과정은 증가폭이 문턱치 이하로 떨어질 때까지 최대 로그 확률을 갖는 노드 분할을 반복한다)

임의의 S를 HMM 상태들의 집합이라 하고 L(S)를 S의 log likelihood라 할때 집합 S에 있는 모든 상태들이 묶였다는 가정 하에 훈련 데이터의 프레임들의 집합인 F에서 생성된다고 하면, 공통 평균 $\mu(S)$과 분산 $\Sigma(S)$를 공유하게 된다.(단 전이 확률은 무시된다) 또한 묶인 상태들의 상태 별 프레임의 정렬이 바뀌지 않는다고 가정하면 L(S)에 대해 다음과 같은 근사식을 쓸 수 있다.

$$L(S) = \sum_{f \in F} \sum_{s \in S} \log(\Pr(o_f; \mu(S), \Sigma(S))) * G \quad (1)$$

(단, $G = \gamma_s(o_f)$)

G : 상태 s에 의해 생성되는 사후 확률

o_f : 관측 프레임

만일 출력 확률 밀도 함수가 가우시안이면 식 (2)와 같이 쓸 수 있다.

$$L(S) = -\frac{1}{2}(\log[(2\pi)_n |\Sigma(S)|] + n) * K \quad (2)$$

(단 $K = \sum_{s \in S} \sum_{f \in F} \gamma_s(o_f)$)

n은 데이터의 차수를 나타낸다. 그러므로 전체 데이터 집합의 로그 확률은 상태들의 분산 Σ(S) 과 전체 상태 점유인 K 만으로 계산할 수 있게 된다. 상태들의 평균과 분산으로부터 $-\frac{1}{2}(\log[(2\pi)_n |\Sigma(S)|] + n)$을 계산할 수 있으며, 상태 점유 수는 이전 단계의 Baum-Welch 재추정하는 동안 저장될 수 있다. 질의어 q에 의해 두 하위 집합 $S_y(q)$와 $S_n(q)$를 나누는 상태들 S와 함께 주어진 노드일 경우, 그 노드는 다음 식 (3) 을 최대화하는 질의 q*를 사용하여 분할한다.

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

식(3)의 $L(S_y(q^*))$ 는 음성학적 질문 q 에 대한 yes로 응답한 node 의 로그유사도(log likelihood), $L(S_n(q^*))$ 는 no로 응답한 node의 로그 유사도이고, $L(S)$ 는 분할하기전의 파티션 전체의 로그유사도이다. ΔL_q 가 최대가 되는 지점에서 질문 과 분할노드를 선택한다. 여기에서 기준이 되는 threshold를 적용하여 로그유사도의 증가량이 threshold 보다 높은 경우에만 분할을 하게 된다.

Ⅳ. 제안한 혼합 가우시안 군집화

1. 가우시안 거리측정 및 합성

거리 측정의 목적은 가장 비슷한 것을 찾아내는 것을 의미한다. 출력 확률 분포를 연속 확률 밀도로 갖는 가우시안의 경우, Euclidean 거리 측정법[10]과 오류율 측정에 기반을 두고 있는 Bhattacharyya 거리 측정법[11]이 있다. 사용한 거리 측정법은 식 (4)와 같다.

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

n은 데이터의 차수를 나타내고, μ_{sk}와 σ_{sk}는 상태 s (i 혹은 j)의 가우시안 분포의 k 번째 평균과 분산이다. Bhttacharyya 거리는 두 가우시안 사이의 겹치는 부분을 측정한다. 거리 범위는 0에서 ∞까지의 값을 가지며 각각 가우시안과 완전히 겹치거나 전혀 겹치지 않음을 나타낸다. 실제로 가우시안은 같은 공간 내에 있으므로 항상 가우시안 간의 겹치는 부분이 조금이라도 있게 되며 거리 측정은 절대 ∞가 되지 않는다. Bhattacharyya 거리 측정의 자세한 식은 식 (5)와 같다.

$$B_{distance}=\frac{1}{8}\left(\vec{\mu}_1-\vec{\mu}_2\right)^T*\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}$$

$$*\left(\vec{\mu}_1-\vec{\mu}_2\right)+\frac{1}{2}\ln\frac{\left|\left(\frac{\Sigma_1+\Sigma_2}{2}\right)\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \quad (5)$$

μ_1과 μ_2은 각 분포의 평균이며 Σ_1과 Σ_2은 공분산이다. Bhattacharyya 거리는 비슷한 가중치를 갖는 가우시안들이 조합되기가 더 쉽기 때문에 축적된(scaled) 형태로 구하게 된다. 이것은 연속적으로 흡수하는 이웃 가우시안들로부터 공분산이 큰 단일 가우시안이 생겨나고, 외부의 것이 흡수되지 않는 것을 막아준다. 가중치 크기(weighting scalar) B_{scale}는 가우시안의 가중치 w_1, w_2의 함수이다.

$$B_{scale}=\sqrt{\frac{w_1^2+w_2^2}{2w_1w_2}} \quad (6)$$

B_{scale}는 $w_1\rightarrow w_2$일 때 $B_{scale}\rightarrow 1$이 된다. 반대로 $w_1\gg w_2$이거나 $w_1\ll w_2$일 때 $B_{scale}\rightarrow\infty$가 된다. 이것은 가우시안 간의 가중치 차이를 최소화하기 위함이다.

$$B_{SD}=B_{scale}B_{distance} \quad (7)$$

가우시안의 각 쌍 간의 거리 B_{SD}를 계산한 뒤에 가장 낮은 B_{SD}를 갖는 쌍이 새로운 가우시안을 생성하기 위해 합성(combine)이 된다. 새로운 가우시안의 파라미터는 가장 낮은 B_{SD}를 갖는 쌍의 가우시안으로부터 생성된다. 새로운 가우시안의 가중치 w_{new}는 두 가우시안의 합과 같다.

$$w_{*w}=w_1+w_2 \quad (8)$$

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

$$\mu_{*w}=\frac{w_1\mu_1+w_2\mu_2}{w_1+w_2} \quad (9)$$

그리고 각 차원을 위해 새로운 분산 σ_{new}은 수정된 분산의 평균의 가중치된 합이다. σ_1과 σ_2를 두 가우시안의 공분산이라 하고 μ_{new}를 식(9) 와 같이 정의하면 새로운 분산 σ_{new}은 다음과 같다.

$$\sigma_{*w}=w_1(\sigma_1+(\mu_1-\mu_{*w})^2)+w_2(\sigma_2+(\mu_2-\mu_{new})^2) \quad (10)$$

그림 2는 가우시안 군집화 과정을 보여주고 있다. 먼저 가우시안들을 메모리에 로딩 하고 ,가장 가까운 두 개의 가우시안을 찾는 과정을 반복한 후에 수렴하는지에 따라 계속할 것인지 종료할 것이지를 결정하게 된다. 가장 가까운 두 개의 가우시안을 찾는 과정은 먼저 가우시안 집합에 가우시안이 있는지를 검사하고 남아 있으면 가장 가까운 두 개의 가우시안을 찾는다. 두 개의 가우시안을 찾으면 합성 목록에 추가한다. 바로 합성하지 않고 목록에 추가하는 이유는 이미 가장 거리가 가까운 가우시안으로 판명이 난 경우 다른 것과의 비교를 피하기 위해서이다. 수렴의 조건은 일정 비율로 줄어들었는지에 대한 여부를 따진다. 이것은 가장 가까운 두 개의 가우시안을 찾는 과정이나 가우시안 합성 과정에서 일어날 수 있다.

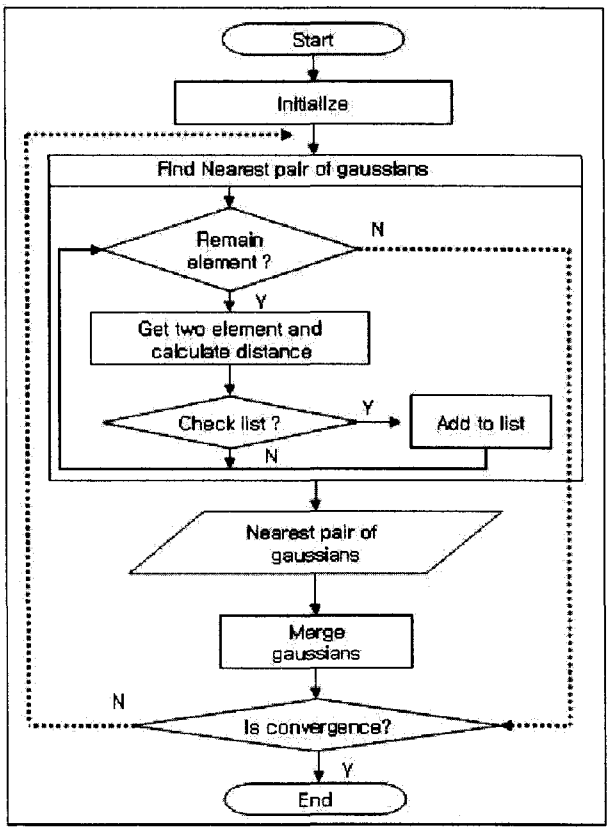

그림 2 가우시안 군집화 과정
Fig. 2. Clustering procedure of Gaussians.

V. 실험 및 고찰

1. 훈련 데이터베이스 분석

실험에 사용한 음성 데이터베이스는 증권상장 회사명(삼성전자,엘지전자,...,교보증권 등) 1,680개 단어이고, 인식시스템은 1680개의 단어만을 인식하는 시스템을 구성하였다. 음향모델을 위한 신호 및 특징 유형은 표2에 보여주고 있으며,CHMM (continuous hidden makov model)의 3상태 Left-to-Right 모델의 원형(prototype)을 취한다. 증권상장 회사명 데이터베이스는 본 연구실에서 취득한 20-30대 남,녀 화가가 발성한 음성으로 총 3 set 로 되어 있으며, 각 set는 560개씩의 증권상장 회사명 어휘로 이루어진다. 전체 데이터베이스는 총 127명분의 음성이고, 훈련에 참가한 데이터베이스는 116명분의 음성이며, 테스트에 참가한 데이터베이스는 학습에 참가하지 않은 11명분의 음성 을 사용하였다. 훈련은 HTK(Hidden markov models toolKit)를 이용하여 수행하였고, 인식평가는 본 연구실에서 자체 제작한 인식기를 사용하였다.

표 2. 데이터베이스 신호 처리
Table 2. Signal processing of database.

설정		설정값	
Environment		조용한 사무실	
Sampling rate, Qantization		8khz,16bit	
Preemphasis coefficient		0.97	
Window	Shape	Hamming	
	Size	25 ms	
	Shift step	10 ms	
Feature	MFCC	26	12
	Normalized energy		1
	1st order derivatives		13

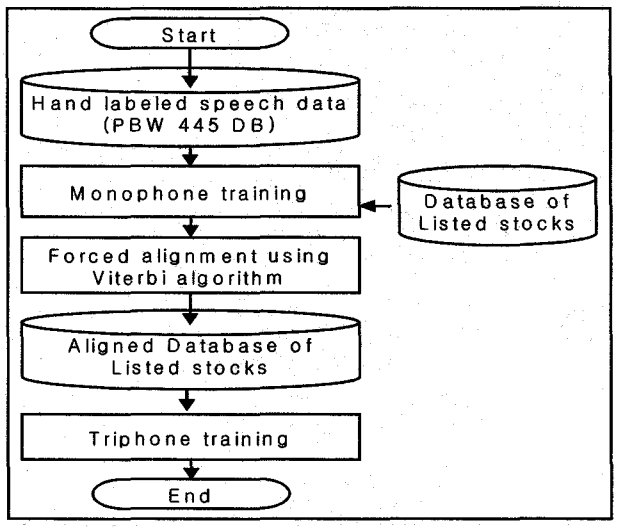

그림 3. 트라이폰 모델링 과정
Fig. 3 Procedure of triphone modeling.

2. 음향모델링

음향 모델생성 과정은 그림 3과 같다. 트라이폰을 정밀하게 모델링하기 위하여 수동 레이블링된 데이터로

표 3. 한국어 변이음 분류표
Table 3. Korean allophonic variations

Classification				Phone
Voiced				ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅣ ㅐ ㅒ ㅔ ㅖ ㅚ ㅙ ㅞ ㅟ ㅢ ㅘ ㅝ ㄴ ㄹ ㅇ ㅇ
Vowel				ㅏ ㅑ ㅓ ㅕ ㅗ ㅛ ㅜ ㅠ ㅡ ㅣ ㅐ ㅒ ㅔ ㅖ ㅚ ㅙ ㅞ ㅟ ㅢ ㅘ ㅝ
vowel	Position of tongue	Fornt	unround	ㅣ ㅐ ㅔ
			Round	ㅚ ㅙ ㅞ ㅟ
		Mid	Unroun	ㅏ ㅓ
		Back	Unround	ㅓ ㅡ
			Round	ㅜ ㅗ
	Width of tongue	Narrow		ㅣ ㅜ ㅗ ㅠ
		Half-wide		ㅖ ㅒ ㅗ ㅕ ㅖ ㅝ ㅟ ㅕ ㅡ
		Wide		ㅏ ㅑ ㅘ
	Height of tongue	High		ㅣ ㅜ ㅟ ㅡ
		Mid high		ㅖ ㅗ ㅚ ㅞ ㅙ ㅕ
		Mid low		ㅒ
		Low		ㅏ ㅓ
	Position of strictur e	Palatal		ㅣ ㅒ ㅖ ㅗ ㅚ ㅞ ㅙ ㅟ
		Velar		ㅜ ㅡ
		Pharynx		ㅏ ㅗ ㅕ
		Close		ㅏ ㅣ ㅜ ㅟ ㅡ
	Apertur e of strictur e	Semi close		ㅖ ㅚ ㅞ ㅙ ㅕ
		Open		ㅒ ㅗ

Classification				Phone
Consona nt	Place of articulati on	Bilabial		ㅏ ㅟ ㅂ ㅃ ㅍ ㅁ
		Dental		ㄷ ㄸ ㅌ ㄴ ㅅ ㅆ ㄹ
		Palatal		ㅑ ㅈ ㅊ ㅉ
		Velar		ㅏ ㅟ ㅢ ㄱ ㄲ ㅋ ㅇ
		Glottal		ㅎ
	Manner of articulati on	Plosive	Unaspirated lenis	ㅂ ㄷ ㄱ
			Aspirated fortis	ㅍ ㅌ ㅋ
			Unaspirated fortis	ㅃ ㄸ ㄲ
		Affricat e	Unaspirated lenis	ㅈ
			Aspirated fortis	ㅉ
			Unaspirated fortis	ㅊ
		Fricativ e	Unaspirated lenis	ㅅ ㅎ
			Unaspirated fortis	ㅆ ㅎ
	Nasal			ㅁ ㄴ ㅇ
	Liquid			ㄹ
	Semi-vowel			ㅑ ㅘ ㅟ ㅢ
	Silence			sil

모노폰을 훈련하고, 이 모노폰으로 훈련하고자 하는 증권상장회사명 데이터베이스를 비터비 알고리즘을 이용하여 음소정렬 한 뒤에 트라이폰 모델을 생성하기 위해 훈련하였다. 이런 과정을 통해 얻어진 정렬된 데이터베이스는 수동 레이블링한 효과를 가지게 되고, 레이블 및 세그먼트 정보를 기반으로 모델링 한 경우 훈련된 모델이 빠르게 수렴하게 되어 훈련 시간을 단축시키고 ,보다 정밀한 모델링을 가능하게 한다. 그림 3에서 생성된 트라이폰을 기반으로 하여 상태공유 모델을 생성하는데 이 때 식(2)에서 필요한 정보를 추출해 낼 수 있다. 정보의 형태는 모델 이름, 모델 출현 회수, 각 상태에서의 관측 확률 값을 나타낸다. 질의어 집합은 표3과 같이 한국어 변이음 분류표를 기반으로 하여 작성되었다.[12] 본 논문에 사용한 Question Set 은 총 7개를 사용하였다. Question Set은 혀의 위치(Position of tongue), 혀의폭(Width of tongue), 구속점(Position of stricture), 구속틈(Aperture of stricture), 조음기관(Place of articulation),조음방법(Manner of articulation)로 분류하였으며, 완성된 Question node 는 Left, Right를 합해 총 150개 이고, Question 형태는 다음과 같다.

QS "R_Vowel"{ *+IY, *+E,..., *+JU, *+JO }
QS "L_Front_Vowel"{ IY-*, E-* }
QS "L_Nasal"{ M-*, N-*, NX-* }

3. 실험 및 평가

상태공유 모델을 생성하기 위해서는 최소 데이터 양이 확보되야 하고, 실험을 통한 기준보다 적은 데이터를 갖는 모델이나 상태에 대해서는 일반적으로 질의어 집합을 통해 통합되어 진다.[1] 또 log likelihood를 구하여 공유 여부를 결정하게 되는데 이는 인식률이 최대인 지점으로 구할 수 있다. 본 논문에서 사용된 데이터베이스에서 출현한 모델 종류의 수와 모델 출현 회수를 보면 그림4 와 같다.

전체 모델의 수는 3,729개이고, 그 중 100개 이하의 출현 회수를 보이는 모델 종류의 수는 2,125개이다. 모델이 통합될수록 데이터의 양은 많아지겠지만 정밀한 모델링이 가능하지 않게 된다. 그림4 를 보면 40번 정도 출현한 모델이 600개 정도이고 20번 정도 출현한 모델은 250개로 나타났다. 상태공유는 로그 유사도를 중심으로 이루어지기 때문에 실험에서는 로그 유사도가 제일 적게나올 수 있다고 판단되는 모델을 택하였다. 실험 결과 10번 이하의 출현 회수를 갖는 모델들로 결

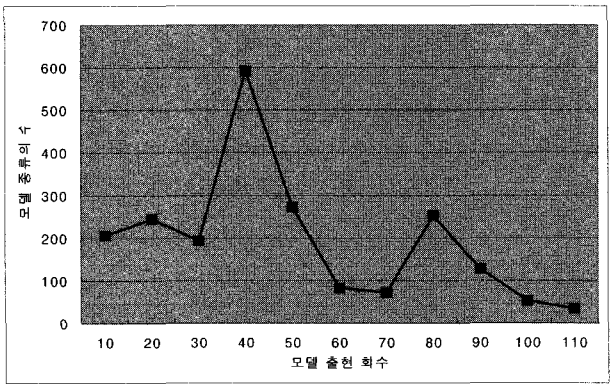

그림 4. 모델 종류의 수와 모델 출현 회수
Fig. 4 The number of kinds of models and the number of models.

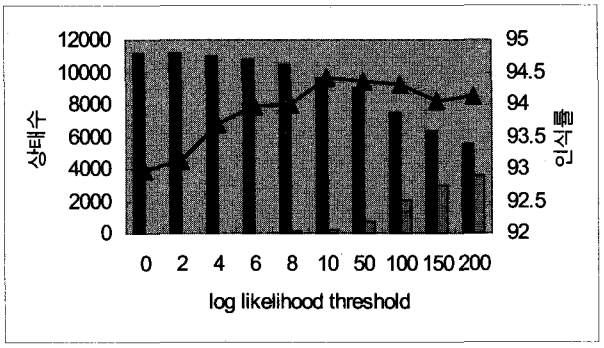

그림 5. l로그 유사도 문턱치 에 따른 상태의 수와 인식률
Fig. 5 The number of states and the recognition ratio according to the log likelihood.

정하였다. 이를 바탕으로 로그 유사도 에 따른 인식률과 상태공유의 실험결과는 그림5 와 표4에 나타내었다.

그림 5 에서 왼쪽 막대는 전체 상태의 수이고, 오른쪽 막대는 공유된 상태의 수이다. 꺾은선은 각각의 인식률을 나타낸다. 로그유사도 문턱치 는 트리의 노드 분할을 결정하기 위한 기준(criterion)으로서 식 3에서와 같이 ΔL_q 가 최대가 되는 지점이다. 실험결과 문턱치 10 이하에서는 상태공유가 거의 이루어지지 않았고 인식률의 변화도 미비하였다. 그러나 문턱치 10 이상에서 상태공유가 눈에 띄게 일어나고, 상태공유 트라이폰 모델의 인식률이 높아짐을 확인할 수 있었다. 최대가 되는 지점은 로그유사도의 문턱치를 10으로 한 지점이었고, 인식률은 로그유사도 문턱치가 높아짐에 따라 감소하다 200이 되는 지점에서 약간 올랐으나 그 뒤 역시 계속 낮아지는 형태를 보였다.

위와 같은 기본 공유 모델 실험을 기반으로 하여 혼합 가우시안의 수를 늘려감에 따라 공유되지 않은 트라이폰 모델과 비교한 것이 표 5 이다. MG#은 상태별 혼

표 4. 로그 유사도 문턱치 에 따른 상태 변화 및 인식률변화
Table 4. state and recognition rate change by log likelihood threshold.

Log likelihood threshold	상태수	공유유상태수	인식률(%)
0	11,187	0	92.97
2	11,170	23	93.12
4	11,018	27	93.68
6	10,805	50	93.98
8	10,456	105	94.00
10	**9,612**	**213**	**94.39**
50	9,035	754	94.35
100	7,488	2031	94.31
150	6,317	2980	94.04
200	5,505	3585	94.13

표 5. 트라이폰 모델과 상태공유 트라이폰 모델
Table 5. Triphone vs. tied state triphone model.

모델 종류		모델수	상태수	혼합 가우시안수	인식률(%)
triphone		3,729	3,729 * 3 = 11,187	11,187	92.9
tied state triphone	MG1	3,729	5,019	5,019	94.3
	MG2	3,729	5,019	11,952	95.6
	MG3	3,729	5,019	17,925	96.7
	MG4	**3,729**	**5,019**	**23,898**	**97.2**

표 6. 상태공유 모델과 혼합 가우시안 군집화모델
Table 6. State Tying model vs. clustering model of mixture Gaussians.

모델 종류	혼합 가우시안수	감소 비율(%)	인식률(%)
Tied state triphone	23,898	0	97.2
Euclidean	23,660	1.0	96.9
Bhattacharyya	23,660	1.0	97.2

합 가우시안의 수를 나타낸다. 트라이폰 모델인 경우 혼합 가우시안의 수를 실험적으로 늘릴 수 없었으며 이는 각 모델에 필요한 데이터 수가 충분하지 않음을 나타내고 있다. 상태공유 모델에서는 혼합 가우시안의 수를 네 개까지 늘릴 수 있었고, 이것이 훈련 데이터에서 일괄적으로 늘릴 수 있는 혼합 가우시안의 수였으며 인식률 역시 최대가 되는 지점이었다.

상태를 공유한 경우 공유된 상태의 수가 5,019개로 트라이폰의 상태수보다 44.86 % 줄어들었음을 확인할 수 있었다. 혼합 가우시안의 증가율은 MG1에서 MG2로 늘릴 때 38.1 %, MG2에서 MG3로 늘릴 때는 49.9 %, MG3에서 MG4로 늘릴 때는 33.3 % 씩 증가하였다. 이 때 각 인식률 증가는 각각 1.3 %, 1.1 %, 0.5 % 씩 증가한 결과를 보였고, 혼합 가우시안을 늘릴 때마다 그 양은 두 배씩 증가하였으나 각각의 인식률의 변화량은 크게 늘지 않음을 알 수 있었다. 이를 근거로 본 논문에서는 인식률을 유지하면서 증가하는 혼합 가우시안의 양을 줄이기 위하여 군집화 기법을 이용하였다. 상태공유 모델과 혼합 가우시안을 통해 군집화한 모델의 비교는 표6 과 같다. 상태공유 모델은 표5 에서 성능이 가장 좋은 MG4를 선택하였고, 혼합 가우시안 군집화 시에 수렴 조건을 혼합가우시안 감소 비율 1.0 % 지점으로 지정하여 인식률을 유지하였다.

감소 비율을 1.0%로 지정한 것은 그 이상에서는 인식률이 현저하게 떨어짐을 보였기 때문이다. 무엇보다도 인식률의 감소 없이 혼합 가우시안의 수를 줄이는 것이 목적이기 때문에 인식률이 감소하지 않는 범위 내에서 실험을 수행하였다. 표6 의 결과를 보면 Euclidean 보다 Bhattacharyya의 성능이 동일한 감소 비율에서 0.3 % 높은 인식률을 보임으로써 더 나은 측정 기법임을 확인할 수 있었다. 하지만 그 성능 차이가 크지 않음을 알 수 있다. 군집화 을 통해 전체 혼합 가우시안의 수를 줄이고자 하였으나 실험 결과 1.0 %의 감소 이후로는 크게 인식률이 떨어졌다.

4. 고찰

본 실험에서는 최적화된 상태공유 모델을 생성하기 위해 혼합 가우시안 군집화를 수행 하였다. 먼저 최적화된 상태공유 모델을 생성하기 위하여 데이터의 최소 출현 회수를 적절하게 정하기 위하여 각 모델의 출현 회수와 모델 종류의 수를 조사하였다. 모델을 생성하기 위하여 데이터를 통합하였을 때 그 통합의 정도가 크게 영향을 미치지 못하도록 출현 빈도가 적은 수의 모델을 선정하였다. 실험에서는 10회 이하로 출현한 모델의 데이터만을 선택 하였고, 이 조건에 해당되는 모델들만을 통합하도록 하였으며, log likelihood의 문턱치 선정은

인식률이 가장 좋은 지점을 최적의 지점이라고 판단하여 이를 이용하였다. 혼합 가우시안의 수를 증가시킬수록 원래의 트라이폰 모델보다 향상된 인식률을 갖는 모델을 생성할 수 있었다. 트라이폰 모델에서의 혼합 가우시안 수의 증가는 데이터 부족으로 수행할 수 없었고, 상태공유 모델에서는 네 개까지만 그 수를 증가시킬 수 있었다. 혼합 가우시안 군집화를 통해 가우시안 수를 줄일 수 있다는 점을 근거로 하여 실험을 하였으나 줄어든 혼합 가우시안의 수를 전체 혼합 가우시안의 1.0 %에서 인식률을 유지하면서 감소시킬 수 있었다. 계산량 감소의 효과를 검증할 만큼 그 수를 많이 줄일 수 없었다. 하지만 상태공유 모델의 혼합 가우시안 수를 줄일 수 있음을 실험을 통하여 확인 하였다.

VI. 결론 및 향후 연구

본 논문은 공유 모델링의 대표적인 방법인 결정트리 기반 상태공유 모델을 기반으로 그 출력 확률 분포의 혼합 가우시안 수를 줄임으로써 모델을 최적화하고자 하였다. 이를 위한 실험에서는 인식률을 최대로 하는 지점을 구하기 위하여 그 문턱치를 변화해 가면서 측정하였고, 실험 결과 문턱치가 커질수록 상태의 공유는 많아졌지만 인식률은 떨어졌음을 확인할 수 있었다. 혼합 가우시안의 군집화는 Euclidean 과 Bhattacharyya 거리 측정 방법을 비교하여 실험한 결과 후자의 측정 방법이 더 나은 성능을 보였다. Bhattacharyya 방법은 97.2 %의 인식률을 유지하면서 전체 혼합 가우시안 수의 비율을 1.0 %로 감소시켰다. 계산량 감소의 효과를 검증할 만큼 그 수를 많이 줄일 수는 없었다. 이의 원인은 상태공유 과정에서 상태가 많이 공유되어 줄어든 수가 많지 않은 것이라 판단된다. 그러나 인식률은 유지하고 상태공유 모델의 혼합 가우시안 수를 줄일 수 있음을 보여 음향모델을 최적화할 수 있음을 실험을 통하여 확인하였다. 향후 공인된 다양한 음성데이터 베이스를 가지고 본 논문에서 제안한 음향모델 최적화 방법으로 인식률이 최대인 점을 유지하면서 가우시안 수를 줄이는 실험을 수행 할 것이다.

참 고 문 헌

[1] S. Young, D. Kershaw, J. Odell, D. Ollason, Valtcher, P. Woodland, "The HTK Book, Cambridge University Engineering Department, 2002.

[2] L. R. Rabiner, B.H. Juang, "Fundamentals of speech recognition", Prentice Hall, New Jersey, chap. 6, 1993.

[3] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE , Volume: 77 Issue: 2 , pp. 257 -286, Feb. 1989.

[4] S. Takahashi. S. Sagayama, "Four-level tied-structure for efficient representation of acoustic modeling", ICASSP-95, International Conference on , Vol.: 1 , pp. 520 -523, May 1995.

[5] K. F. Lee,"Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition",Acoustics, Speech, and Signal Processing, IEEE Transactions on , Volume: 38 Issue: 4 pp. 599 -609, Apr. 1990.

[6] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree based state tying forhigh accuracy modeling," in ARPA Workshop Human Language Technology,Princeton, NJ, pp. 286-291, Mar. 1994.

[7] J. R. Bellegarda, D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition", Acoustics, Speech, and Signal Processing, IEEE Transactions on , Volume: 38 Issue: 12 pp. 2033 -2045, Dec. 1990.

[8] W. Reichl, Wu Chou, "Robust decision tree state tying for continuous speech recognition", Speech and Audio Processing, IEEE Transactions on , Volume: 8 Issue: 5 pp. 555 -566, Sep. 2000.

[9] A. Kannan, M. Ostendorf, J.R. Rohlicek, "Maximum likelihood clustering of Gaussians for speech recognition", Speech and Audio Processing, IEEE Transactions on , Volume: 2 Issue: 3 pp.453 -455, Jul. 1994.

[10] J. J. Odell, "The use of context in large vocabulary speech recognition", PhD's Dissertation. University of Cambridge. 1995.

[11] K. Fukunaga, "Introduction to statistical pattern recognition", Morgan Kaufman, San Francisco, p.97-99, 1990.

[12] 오세진, 황철준, 김범국, 정호열, 정현열, "결정트리 상태 클러스터링에 의한 HM-net 구조결정 알고리즘을 이용한 음성인식에 관한 연구", 한국음향학회지 제 21권 제2호, pp. 199-210, 2002.

[13] J. Takami, S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling", ICASSP-92., p. 573 -576, Mar., 1992.

저 자 소 개

안 태 옥(정회원)
제 41권 SP편 제 3호 참조