동적 다중 그룹 혼합 가중치를 이용한 한국어 음성 인식의 성능향상

황기찬[O] 김종광 김진수 이정현
인 하 대 학 교 전 자 계 산 공 학 과
[gchwang, bluemir, kjspace]@nlsun.inha.ac.kr jhlee@inha.ac.kr

Improvement in Korean Speech Recognition using Dynamic Multi-Group Mixture Weight

Gi-Chan Hwang[O] Jong-Kwang Kim Jung-Hyun Lee
*Dept. of Computer Science & Engineering, Inha University

요 약

본 논문은 CDHMM(Continuous Density Hidden Markov Model)의 훈련하는 방법을 동적 다중 그룹 혼합 가중치(Dynamic Mutli-Group mixture weight)을 이용하여 재구성하는 방법을 제안한다. 음성은 Hidden 상태열에 의하여 특성화되고, 각 상태는 가중된 혼합 가우시안 밀도 함수에 의해 표현된다. 음성신호를 더욱더 정확하게 계산하려면 각 상태를 위한 가우시안 함수를 더욱더 많이 사용해야 하며 이것은 많은 계산량이 요구된다. 이러한 문제는 가우시안 분포 확률의 통계적인 평균을 이용하면 계산량을 줄일 수 있다. 그러나 이러한 기존의 방법들은 다양한 화자의 발화속도와 가중치의 적용이 적합하지 못하여 인식률을 저하시키는 단점을 가지고 있다. 이 문제를 다양한 화자의 발화속도에 적합하도록 화자의 화자의 발화속도에 따라 동적으로 5개의 그룹으로 구성하고 동적 다중 그룹 혼합 가중치를 적용하여 CDHMM 파라미터를 재 구성함으로써 8.5%의 인식율이 증가되었다.

1. 서론

CDHMM(Continuous Density HMM)은 음성인식에 성공적으로 적용이 되어져왔다. HMM은 초기 밀도,상태 천이 확률,관측 밀도로 특성화된다. HMM의 학습에서 관측밀도가 중요하다. HMM의 최대 성능향상을 위하여 CDHMM을 사용하고 CDHMM에서 각 상태들은 가중치가 적용된 혼합 가우시안 밀도 함수에 의하여 특성화가 된다. 음성인식기의 성능개선을 하려면 일반적으로 전향 알고리즘으로 더 많은 가우시안 분포를 HMM을 이용하여 훈련을 해야 한다. 즉, 음성인식과 훈련의 시간이 적은 가우시안 분포의 HMM을 훈련할 때 보다 더 많은 시간을 소비하게 되는 문제점이 있다. 이러한 문제점을 해결하기 위하여 관측열에 의하여 훈련되는 데이터를 최적화하기 위한 수많은 학습 알고리즘들이 개발되어져 왔다. 본 논문은 CDHMM 파라미터를 DMGMW (Dynamic Multi-group Mixture Weight)으로 재구성하여 최적화함으로써 인식율을 증가시키는 방법을 제안한다.

2. 관련연구

2.1 Viterbi decoding

Viterbi decoding은 최적 상태관측열(state sequence)과 최적 상태관측열이 나올 확률을 구하는 과정이다[1].

훈련부분에서는 최적 상태관측열을 사용하고 인식에서는 최적 상태 관측열에서 나올 확률을사용한다. 비터비 알고리즘 (Viterbi algorithm)은 $P(q|O,\lambda)$를 최대화하는 한개의 상태관측열을 구하는 방법이다.

$$\delta_t(i) = max_{q_1,q_2,\ldots,q_{t-1}} P[q_1 q_1 \cdots q_{t-1} = i, o_1 o_2 o_t | \lambda] \quad 식(1)$$

이 식은 한개의 상태관측열를 따라 시간 t일 때 상태 s_i에 있을 가장 높은 확률값을 나타낸다. Recursive한 방법에 의해 다음 시간 t+1에서의 확률값은 다음식과 같이 표시 된다.

$$\delta_{t+1}(j) = [max_i \delta_t(i) a_{ij}] b_j(o_{t+1}) \quad 식(2)$$

최적 상태 관측열은 위의 식을 이용하여 다음과 같이 구한다.

(1) 초기화

$\delta_1(i) = \pi_i b_i(o_1)$, $1 \leq i \leq N$, $\phi_1(1) = 0$

(2) 반복계산

$\delta_t(j) = max_{1 \leq i \leq N}[\delta_{t-1}(i) a_{ij}] b_j(o_t)$, $\begin{cases} 2 \leq t \leq T \\ 1 \leq j \leq N \end{cases}$

$\phi_t(j) = arg\ max_{1 \leq i \leq N}[\delta_{t-1}(i) a_{ij}]$, $\begin{cases} 2 \leq t \leq T \\ 1 \leq j \leq N \end{cases}$

(3) 종료

$P^* = max_{1 \leq i \leq N}[\delta_T(i)]$, $q_T^* = arg\,max_{1 \leq i \leq N}[\delta_T(i)]$

(4) 최적 상태 관측열 찾기

$q_T^* = \phi(q_{t+1}^*)$, $t = T-1, T-2, \ldots, 1$

그림1. Viterbi decoding의 처리 단계

2.2 CDHMM

관측심볼확률 $b_j(o)$을 CDHMM으로 표현하면 아래와 같다[2].

$$b_j(o) = \sum_{k=1}^{M} c_{jk} b_{jk}(X)$$

$$= \sum_{k=1}^{M} c_{jk} N(o, \mu_{jk}, \sum_{jk}) \quad 1 \leq j \leq N \qquad \text{식(3)}$$

가우시안(Gaussian) 분포에 가중치를 준형태로 나타낸다.
여기서 c_{jk}는 가중치(Weighting) 상수, μ_{jk}는 가우시안 분포(Gaussian distribution)의 평균, $\sum_{jk}$는 직교행렬(Covariance matrix), M은 혼합분포(Mixture distribution)의 수를 나타낸다.

3. 연구 구성

본 논문에서 제안한 DMGMW을 이용하여 훈련데이타 셋을 구성하는 과정은 다음과 같다.

3.1 훈련 시스템 구성도

아래의 그림 2에서 DMGMW의 그룹 가중치를 계산하여 초기값을 획득하고 화자의 발화시간에 따라 그룹을 나누어 훈련한다. 3.4의 인식부분에서는 입력되는 화자의 발화 시간에 따라 그룹을 선택하고 이 그룹의 훈련 데이터를 이용하여 인식을 한다.

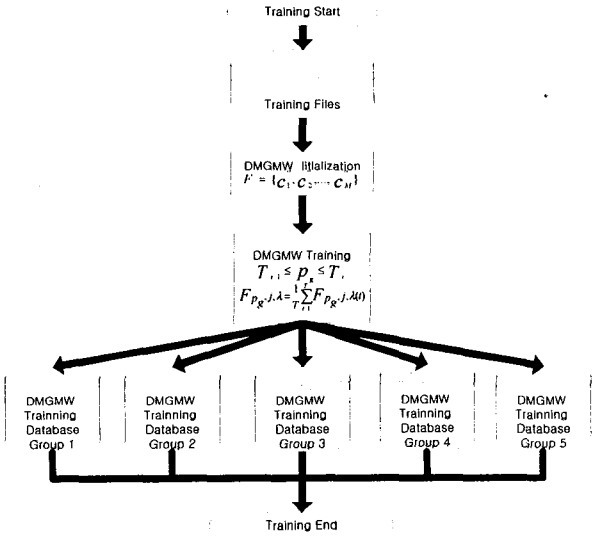

그림 2. 훈련시스템 구성도

3.2 DMGMW의 초기 파라미터

CDHMM에서 사용되는 식은 식(4)와 같다[2]. 초기 파라미터들은 최종의 CDHMM의 품질에 영향을 주는 가장 중요한 요소이다. 그러므로 DMGMW의 초기 파라미터를 알맞게 구성하는 것이 중요하다. 식(4)를 이용하여 초기파라미터를 구성한다.

$$\xi_t(j,k) = f(s_t = j, k_t = k, | X, \lambda) \quad 1 < t \leq T \qquad \text{식(4)}$$

$\xi_t(j,k)$는 주어진 모델 λ와 관측열 X에서 시간 t에서 k번째의혼합 요소와 상태가 j일 때 모든 가능한 전이확률을 나타낸다. 시간 t에서 관측벡터 X_t과 모델 λ를 식(5)~식(6)와 같이 표현한다[2].

$$\eta_t(j) = \sum_{k=1}^{M} \xi_t(j.k) \qquad \text{식(5)}$$

$$F_j(X_t) = \left\{ \frac{\xi_t(j,1)}{\eta_t(j)}, \frac{\xi_t(j,2)}{\eta_t(j)}, \ldots, \frac{\xi_t(j,M)}{\eta_t(j)} \right\} \qquad \text{식(6)}$$

M은 혼합 분포의 수를 나타낸다. $F_j(X_t)$는 M차원의 벡터로서 혼합 가중 벡터(Mixture weight vector)라 한다.

CDHMM의 초기 파라미터는 훈련 데이터에서 비터비 디코딩[1]을 사용하여 획득하고 식(6)을 사용하여 혼합 가중 벡터들을 계산한다. 그러나 초기 가중치를 관측할 때 몇 개의 가중치가 0과 1에 동시에 가깝게 나타나는 것을 찾을 수 있다. 이러한 경우에 CDHMM의 성능을 저하시킨다. 이 문제를 해결하기 위하여 식(7)~식(8)에 따라 급속한 변동을 감소시킨다[5].

혼합 가중치 그룹 $F = \{c_1, c_2, c_3, \ldots, c_M\}$는 $\sum_{i=1}^{M} c_i = 1$을 만족해야한다. 식(7)에 의하여 성능을 저하시키는 급속한 변동을 감소시킨다.

$$c_i' = c_i'^{*}\theta + (1 - c_i)^{*}(1 - \theta) \quad 1 \leq i \leq M \qquad \text{식(7)}$$

여기서 $\theta = \frac{\epsilon^{*}(M-1)}{1+\epsilon^{*}(M-2)}$ $\epsilon = 0.5 \sim 0.7$ 식(8)

식(7)에 의하여 변환하면 c_i'가 된다. 여기서 $1 \leq i \leq M$에서 $\sum_{i=1}^{M} c_i' = 1$을 만족하지 못하므로 식(9)와 식(10)을 이용하여 식(7)에 의한 문제를 해결한다[5].

$$C' = \sum_{i} c_i' \qquad \text{식(9)}$$

$$c_i'' = c_i'/C' \quad 1 \leq i \leq M \qquad \text{식(10)}$$

마지막으로 식(9)를 바탕으로 DMGMW의 초기값은 식(11)으로 구성된다[5].

$$F = \{c_1'', c_2'', \ldots, c_M''\} \qquad \text{식(11)}$$

3.3 DMGMW의 훈련

DMGMW의 훈련은 아래와 같이 식으로 발화속도에 따라 5개의 그룹으로 나누었다. 훈련에서 새로운 훈련데이타의 삭제, 추가, 대치에 따른 그룹의 재구성은 발화시간에 따라 재구성을 한다.

$$MT = MAX\ \{W_i(t)\}/g \qquad g = 5 \qquad \text{식(12)}$$

$$T_i = T_{i-1} + MT \qquad \begin{cases} 1 \leq i \leq 5 \\ T_0 = 0 \end{cases} \qquad \text{식(13)}$$

식(12)에서 MT는 전체 단어에서 화자의 발화속도에 따른 최대값을 그룹으로 나눈 각 그룹사이에 발화 시간간격이며 $W_i(t)$는 각 단어의 시간이다. 식(13)에서 T_i는 그룹을 구성하기 위한 그룹간의 시간 간격, g는 그룹의 개수이다. 즉 그룹을 구성하는데 단어의 최대값을 이용한다. 그러나 단어의 최대값이 너무 크거나

변화이 심하면 시스템의 성능에 저하를 가져온다. 이 시스템의 실험에서는 발화속도에 따른 최대값의 변동의 상한·하한을 ± 150msec로 제한을 한다. 식(6)로 관측 벡터 X_t를 모델 λ에 상태 j를 할당하는 처리를 한다. 출력확률을 계산하고 각각의 그룹에 혼합 가중치를 위한 X_t의 혼합 가중치를 계산한다. 그리고 최대 출력 확률의 혼합 가중 벡터를 기록하고 인덱싱을 한다.

식(12)~식(15)에 의하여 가중 벡터를 계산한다. 그리고 EM 알고리즘으로 DMGMW의 훈련파라미터를 만든다[6][7].

$$\hat{F}p_g, j, \lambda = \frac{1}{T}\sum_{t=1}^{T} Fp_g, j, \lambda(t) \qquad 1 \leq g \leq 5 \quad \text{식(14)}$$

$$T_{i-1} \leq p_g \leq T_i \qquad \begin{cases} 1 \leq g \leq 5 \\ 1 \leq i \leq 5 \end{cases} \quad \text{식(15)}$$

3.4 DMGMW의 인식

인식을 위한 DMGMW은 발화속도에 따라 인식그룹을 선택하고 식(16)에 따라 음성인식을 한다.

$$b_j(o_t) = max\sum_{k=1}^{M} c_{p_g k} N(o_t, \mu, \sum) \qquad 1 \leq g \leq 5 \quad \text{식(16)}$$

$b_j(o_t)$는 상태열 j에서 관측 o_t의 관측을 위한 출력확률 , $c_{p,k}$는 화자의 발화속도에 따른 p번째 그룹의 k번째 가중치, N(.)은 가우시안 밀도함수이다.

식(16)에서 가우시안 밀도 함수의 확률을 구하고 각 그룹의 동적 혼합 가중치를 곱하여 출력확률을 계산한다.

4. 실험 및 결과

전처리에 의한 특징 추출은 16kHz의 MFCC 12차 멜켑스트럼 계수(Mel-cepstral coefficient)를 사용하였다. DMGMW의 학습에는 표 1과 같이 20대에서 40대의 50명의 수집된 화자의 훈련 데이터베이스와 4명(남 2, 여 2)의 테스트 데이타를 가지고 실험을 한다.

표 1 훈련 데이타 베이스

일	뉴스	농구권	국초고왕	멀티미디어
이	미용	국조도	정비대원	서울기독교청년회
삼	심회	소규모	한크리드	서울대학교
사	뉴욕	승용차	성냥개비	경영다각화
오	오빠	코코아	민족기업	비영농정책환동
육	위험	병아리	경영소득	금융실명제
칠	적응	장애물	교통사고	광양제철소
팔	계승	다람쥐	로네상스	대동여지도
구	계약	장애인	다품종소량	기계모내기
십	시인	라디오	스케치북	아리스토텔레스
하나	전용	소득과	스크랩북	최대공약수
둘	시험	보따리	네델란드	아르바이트
셋	유령	대법원	자동차용	인하대학교
넷	학습	국세청	국민투표	추모위원회
다섯	외과	프리킥	우리사회	추진위원회
여섯	개관	선거권	기업규범	농촌공업화
일곱	금융	정의선	지역의원	엘리베이터
여덟	실험	나침반	자연언어	한국소비자보호원
아홉	전투	오존층	제품개발	중앙집권화
영	하원	제래식	유기용제	마르크스주의

실험비교를 하기 위하여 기존에 사용된 CDHMM과 DMGMW을 만들어 비교를 하였다. 화자의 발화속도에 따라 5개의 p그룹을 만들어 실험을 하였다. 그리고 표 2과 같이 각 화자를 두 시스템을 이용하여 비교하여 평가하였다.

표 2 CDHMM과 DMGMW의 인식 성능 실험 비교 결과

	CDHMM	DMGMW
화자 1[남]	86.5 %	95.5 %
화자 2[여]	89.7 %	98.8 %
화자 3[남]	87.1 %	96.3 %
화자 4[여]	84.8 %	91.5 %

그림 3은 표 2를 바탕으로 실험의 비교 결과를 나타낸 그림이다. CDHMM은 평균 87.03%를 나타내고 DMGMW는 평균 95.53%의 인식율을 나타내었다. 그림과 같이 전체적으로 두시스템 간에 8.5%의 차이를 나타낸다.

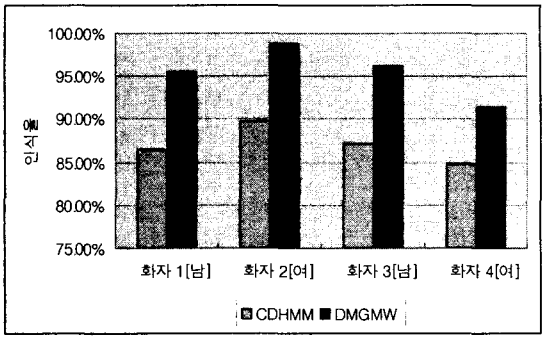

그림 3 CDHMM과 DMGMW의 인식 성능 실험 비교 결과

5. 결론

기존의 CDHMM은 단지 하나의 그룹과 고정된 가중치를 적용하였다. 이것은 다양한 화자 특성에 적합하지 못하여 인식율을 저하시킨다. 이러한 문제점을 DMGMW를 사용하여 발화속도에 따라 5개의 그룹으로 나누어 가중치를 적용함으로써 다양한 화자의 특성에 적합하도록 하였다. 본 논문에 제안된 방법에 의하여 재구성을 함으로써 전체적으로 8.5%의 인식율이 향상되었다. 이 시스템은 그룹을 동적으로 구성하고 이를 데이터베이스를 사용하여 인식을 함으로 인식율이 향상되고 인식시간이 단축이 된다. 그러나, 훈련 데이터 양이 많을수록 동적으로 재구성하는데 많은 시간이 소요가 되는 단점이 있다. 향후 연구 과제로 이 재구성으로 인한 훈련시간의 증가의 문제를 해결이 요구된다.

참고 문헌

[1] Viterbi A. J., "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", IEEE Trans. on Information Theory, IT-13(2), pp.260-269, April 1967.

[2] Huang X.D., Ariki Y., and Jack MA., "Hidden Markov Models for SpeechRecognition", Edinburgh University Press, 1990.

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

[4] F. Jelinek, "Continous speech recognition by statistical methods", Proc. IEEE, Vol. 64, No. 4,pp.532-556, April 1976.

[5] Li Ming and Yu Tiecheng, "Multi-group Mixture Weight HMM", ICSLP, pp.290-292, October 2000.

[6] Baum L.E., Petrie T., Soules G., Weiss N., "A maximum technique occurring in the statistical analysis of probabilistic functions of Markov chains", Ann. Math. Stat., Vol. 41, pp.164-171, 1970.

[7] Baum L.E., "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes", Inequalities, Vol. 3,pp 1-8, 1972.