김일성종합대학학보
(자연과학)
주체103(2014)년 제60권 제7호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 60 No. 7 JUCHE103(2014).

효률적인 HSMM모형파라메터추정알고리듬에 대한 연구

리정철, 김덕호

HSMM에 기초한 음성인식체계들에서 상태지속은 자체이행에 의하여 모형화되는데 이때 상태지속은 지수분포에 따르게 되며 이것은 상태지속확률이 시간에 따라 지수적으로 감소되는 결함을 가진다. 실제적으로 지속은 발성속도, 력점들, 문맥, 음성의 류형(흐린음성 혹은 랑독형)에 따라 고정불변하지 않으며 따라서 음향음운학리론과 음성단위음들사이 음향학적인 련관에서 중요한 역할을 놀고있다.[1]

그리하여 더 일반적인 지속모형을 리용하여 음성에서 가변사건들의 지속과정을 충분히 표현하기 위한 여러가지 모형들(HSMM, Autoregressive(AR)-HMM, ANN(Artificial Neural Networks), Neural Network/HMM, Stochastic Segment model, Segmental Neural Network)이 연구되여왔으며 이에 대응하여 파라메터추정방법들도 갱신되여왔다.

HSMM을 리용하는 경우 파라메터추정법[2]이 제기되였지만 관측렬의 동시확률분포계산식을 계산할 때 매 시간 t에서 이전의 D개의 관측들이 선택되여야 하므로 총체적인 재귀결음수가 D배로 늘어나는것으로 하여 실현하기 힘든 문제가 있었다.

론문에서는 HSMM의 파라메터들을 효률적으로 추정하기 위해 앞방향－뒤방향추정알고리듬을 일반화하였다.

1. HSMM에 기초한 지속모형화와 파라메터추정

1) HMM에 기초한 지속모형화

음성인식에 보편적으로 리용되고있는 HMM에서 상태 i를 지속 τ시간동안 차지할 지속확률은 다음과 같이 표시된다.

$$P(dur=\tau)=(1-a_{ii})a_{ii}^{\tau} \tag{1}$$

이때 상태지속모형은 평균과 분산이 각각 다음과 같이 표시되는 기하분포에 따른다.

$$E(dur)=\frac{1}{1-a_{ii}},\ V(dur)=\frac{a_{ii}}{(1-a_{ii})^2}=E[dur^2]-E[dur]^2$$

식 (1)의 상태지속모형은 단 1개의 자유파라메터에 의하여 결정되며 평균지속이 일단 결정되면 분산은 고정된다. 이와 같이 기하분포는 HMM에 기초한 음성인식에서 음향모형화에 최량이 못된다. 또한 HMM뿐만아니라 다른 복잡한 지속모형에서도 나타나는 문제이지만 단어지속평균은 잘 모형화되지만 단어지속분산은 잘 모형화되지 못한다는것을 실험적으로 보여주었다.

2) HSMM에 기초한 양적인 지속모형화[2]

이제 $S(\in S_T)$는 상태경로, S_T는 마르꼬브사슬에서 길이가 T인 모든 상태경로들의 모임

이라고 하자. 그리고 m번째 상태모임을 s_m, 지속시간을 $\tau_m(1\le\tau_m\le D)$이라고 하면 상태경로는 다음과 같이 표시할수 있다.

$$S=\{(s_m,\ \tau_m)\}_1^M,\ M\le T$$

이때 1계마르꼬브사슬의 성질로부터 다음의 식들이 성립한다.

$$a_{s_{m-1},\ s_m}=\Pr(s_m\,|s_1,\cdots,\ s_{m-1})=P(s_m\,|\,s_{m-1})\equiv a_{m-1\ m}$$

$$d_{s_m}(\tau)=P(\tau_m\,|s_m,\ (\tau_1,\ s_1),\cdots,\ (\tau_{m-1},\ s_{m-1}))=P(\tau_m\,|\,s_m)\equiv d_m(\tau_m)$$

이로부터 $P(S)=P(s_1)P(\tau_1\,|\,s_1)\prod_{m=2}^{M}P(s_m\,|s_{m-1})\Pr(\tau_m\,|s_m)$ 이다.

한편 HSMM상태렬 S와 관측벡토르렬 O사이의 동시우도는 다음과 같이 표시된다.

$$P_\lambda(O,\ S)=\pi_{s_1}d_{s_1}(\tau_1)\prod_{t=1}^{\tau_1}b_{s_1}(o_t)\prod_{m=2}^{M}a_{s_{m-1}s_m}d_{s_m}(\tau_{s_m})\prod_{t'=1}^{\tau_m}b_{s_m}(o_{f_m+t'}) \tag{2}$$

여기서 $f_m=\sum_{n=1}^{m-1}\tau_n$ 이다.

3) HSMM파라메터추정을 위한 효률적인 앞방향-뒤방향(FB)알고리듬

우선 앞방향변수를

$$\alpha_t(m,\ d)=P[o_1^t,(q_t,\ \tau_t)=(s_m,\ d)] \tag{3}$$

로 정의하자. 여기서 $q_t=s_m$은 t번째 시각의 상태, $\tau_t=d$는 상태 q_t에 머무르는 시간이다.

그러면 $(q_t,\ \tau_t)$에로의 이행은 $(q_{t-1},\ \tau_{t-1})=(s_m,\ d+1)$이거나 $(q_{t-1},\ \tau_{t-1})=(s_n,\ 1|n\ne m)$로부터 발생할수 있다. 그러므로 초기조건이

$$\alpha_1(m,\ d)=\pi_m\cdot b_m(o_1)p_m(d) \tag{4}$$

일 때 상태 s_m과 $t>1$에 대하여 식 (3)은 다음과 같이 된다.

$$\alpha_t(m,\ d)=\alpha_{t-1}(m,d+1)b_m(o_t)+\left(\sum_{n\ne m}\alpha_{t-1}(n,\ 1)a_{nm}\right)\cdot b_m(o_t)\,p_m(d),\quad d\ge 1 \tag{5}$$

한편 뒤방향변수를

$$\beta_t(m,\ d)=P[o_{t+1}^T,\ (q_t,\ \tau_t)=(s_m,\ d)] \tag{6}$$

라고 하면 $(q_t,\ \tau_t)=(s_m,\ d)$ 뒤로 올수 있는 가능한 상태들은 $d>1$일 때 다음 상태로는 $(q_{t+1},\ \tau_{t+1})=(s_m,\ d-1)$이 되며 $d=1$일 때에는 $(q_{t+1},\ \tau_{t+1})=(s_n,\ d'|n\ne m,\ d'\ge 1)$이 된다.

여기로부터 초기조건이

$$\beta_T(m,\ d)=1,\quad d\ge 1 \tag{7}$$

이고 상태 s_m, 시간이 $t<T$일 때 식 (6)은 다음과 같이 된다.

$$\beta_t(m,\ d)=b_m(o_{t+1})\beta_{t+1}(m,\ d-1),\quad d>1 \tag{8}$$

$$\beta_t(m,\ 1)=\sum_{n\ne m}a_{mn}b_n(o_{t+1})\left(\sum_{d\ge 1}p_n(d)\beta_{t+1}(n,\ d)\right) \tag{9}$$

우와 같은 고찰로부터 식 (5)의 앞방향변수 $\alpha_t(m,\ d)$를 계산하자면 먼저 모든 m에 대하여

$$\left(\sum_{n\neq m}\alpha_{t-1}(n,\ 1)a_{nm}\right)$$

을 계산하고($O(M^2)$) 모든 m, d에 대하여 $\alpha_t(m,\ d)$를 계산하게 된다.($O(MD)$) 따라서 매 t시각에 앞방향변수를 계산하자면 $O(MD+M^2)$의 회수가 요구된다.

류사하게 뒤방향공식에 대해서도 매 t시각에 $\beta_t(m,\ d)$를 계산하는데 $O(MD+M^2)$만한 회수가 요구된다. 따라서 제안된 알고리듬에서 앞방향－뒤방향변수들의 총 계산회수는 $O((MD+M^2)T)$로서 계산회수가 $O((MD^2+M^2)T)$이던 종전의 추정알고리듬에 비하여 대폭 감소되였다.

2. 평 가 실 험

음향특징벡토르는 평균정규화된 멜케프스트람(MFCC)과 그것의 1계, 2계시간도함수정보를 포함한 39차원 MFCC_0DAZ파라메터이다.

음향모형은 남성, 녀성음성자료를 포함한 80GByte의 음향자료기지를 리용하여 구축되였으며 그것을 조선어련속음성인식체계에 적용하였다. 인식음성자료로서 단어인식용으로 깨끗한 환경에서 록음한 1 500개 단어음성자료와 련속음성인식용으로 서로 다른 마이크로 잡은 일반랑독형문장음성자료 46, 100, 400개를 리용하였다. 매 부류의 검사문장들에 대해서 우리는 발성속도측면에서 볼 때 각각 0.128, 0.140, 0.145s의 평균음절발성길이를 측정하였다.

론문에서는 매 HMM상태의 최대지속허용한계 D는 $\mu+10\cdot\sqrt{\sigma}+0.5$의 웅근수부로 설정하였다. 여기서 μ, σ는 각각 지속분포의 평균과 분산이다.

HSMM의 성능을 단어인식단계에서 분석하기 위하여 700MByte훈련자료를 가지고 훈련된 문맥독립모형(상태당 1개의 가우스분포)을 가지고 1 500개 단어에 대해서 단어인식실험을 진행하였다. 결과 상대적으로 약 2%의 성능이 개선되였다. 실험에 의하면 지속분포의 분산바닥값이 인식성능에 큰 영향을 미치였으며 약 0.4로 하였을 때 가장 성능이 좋았다.

다음으로 문장인식실험에서는 상태당 22개의 혼합가우스분포를 가진 문맥의존HSMM을 리용하여 98.01, 98.18, 97.57%의 인식률을 얻었다. 이것은 기준모형으로 얻은 98.00, 95.94, 96.31%에 비하여 100, 400개 문장에서 뚜렷한 리득이 얻어졌다는것을 보여준다.

0.01, 2.33, 1.30%의 상대적인식률개선과 매 발성자료기지의 특성으로 보아 비교적 느린 문장들에 대해서 HSMM모형에서의 개선이 뚜렷하다고 본다.

맺 는 말

실험을 통하여 음성의 림시적구조를 표현하는데서 HSMM의 인식성능이 HMM에 비하여 우월하다는것을 확증하였으며 모형훈련때 계산비용은 종전의 HMM에 기초한 방법에 비하여 조금 증가되였을뿐이다.

참 고 문 헌

[1] M. M. Hochberg; A Comparison of State Duration Modeling Techniques for Connected Speech Recognition, 12, 1993.

[2] J. Yamagishi et al.; Proc. of ASJ, 3, 225, 2004.

주체103(2014)년 3월 5일 원고접수

HSMM-based Duration Modeling Method in Acoustic Modeling for Speech Recognition

Ri Jong Chol, Kim Tok Ho

In order to model the duration process appropriate to speech production, the cost problems for implementing HSMM (Hidden Semi-Markov Model)-based acoustic modeling is solved by using the efficient estimation algorithm of parameters. We showed the advantage of the HSMMs compared to the HMMs, especially for slow speakers, via our experiments.

Key words: hidden semi-Markov model, duration modeling