隐马尔可夫模型中一种新的帧相关建模方法*

郭 庆 吴文虎 方棣棠

(清华大学计算机科学与技术系语音实验室 北京 100084)

摘要 在使用传统的隐马尔可夫模型(traditional hidden Markov model,简称THMM)刻画现实中的语音时有一个明显的缺点,即THMM不能合适地表征语音信号的时域结构.时域上的相关性被认为对识别非常有用,因为相邻帧间的特征矢量具有很强的相关性.文章提出了一种新的方法,用以把时域的相关性糅合到一个基于传统的隐马尔可夫模型的语音识别系统中.首先,用条件概率的形式处理帧间相关性;然后,用一种非线性的概率近似公式来表征相邻帧之间的相关性.此方法丝毫不增加原来的THMM的空间复杂度,而且也几乎不增加训练和识别阶段的时间复杂度.最后,糅合了帧间相关性的HMM(文章称之为FC(frame correlation) HMM)的首选识别率比原先的THMM提高了6个百分点.

关键词 连续隐马尔可夫模型,帧间相关性,非线性估计,混合高斯密度,联合条件概率密度.

中图法分类号 TP391

在传统的隐马尔可夫模型(traditional hidden Markov model,简称THMM)中,模型在某状态停留一定时间的概率随着时间的增长呈指数下降的趋势,这使得THMM不能合适地表征语音信号的时域结构.为了弥补THMM的这一缺点,人们提出了许多种方法试图将一些额外的信息加入到传统的HMM中去.其中较为典型的方法有:加入状态停留时间、加入高阶特征和利用相邻帧观察矢量之间的相关性等.

在实际的语音识别模型化时,人们采用多种方法来处理帧间相关性.M. Ostendorf[1]等人提出了随机区段模型(stochastic segment model).V. Digalakis[2]等人提出了动态系统模型(dynamical system model).这两种方法均试图直接刻画语音特征的轨迹,尽管它们在合适的轨迹假设下可能会为语音识别抽取出动态信息,但它们均不是基于目前获得广泛应用并取得巨大成功的HMM技术.

在连续隐马尔可夫模型中,C. J. Wellekens[3]对相邻帧间的特征向量定义了一种高斯概率密度函数.P. Kenny[4]等人尝试用线性预测技术来参数化帧间相关性.

在离散隐马尔可夫模型中,Paliwal[5]把一个观察矢量的概率分布构筑于每一对状态和前一观察输出符号之上.这种完全参数化的方法导致了待估计参数数目的急剧增长,对于有限的训练数据,要想获得可靠的参数估计几乎是不可能的.S. Takahashi[6,7]提出了BC(bigram-constrained) HMM,这种模型很好地避免了上述问题.在BC HMM中,一个观察矢量的输出概率同样也依赖于当前状态和前一时刻的观察矢量,但它们是分别进行估计,然后再组合计算得到的.这样就使得BC HMM要估计的参数数目远远少于Paliwal提出的完全参数化方法所要求的数目.BC HMM的一个显著特点即在于,它提供了一种由单个条件的概率分布已知,进而近似估计联合条件概率分布的方法.N. S. Kim[8]提出了一种基于扩展对数池(extended logarithmic pool)的算法,从而更加精确地估计联合条件概率分布.

---

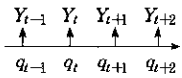
* 本文研究得到国家"九五"攻关项目基金资助.作者**郭庆**,1970年生,博士生,主要研究领域为语音识别与理解.**吴文虎**,1936年生,教授,博士生导师,主要研究领域为人工智能,汉语语音识别与理解.**方棣棠**,1930年生,教授,主要研究领域为信号处理,信息论,语音信号处理,语音识别.

本文通讯联系人:郭庆,北京 100084,清华大学计算机科学与技术系语音实验室

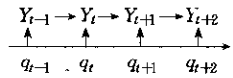
本文1998-04-27收到原稿,1998-06-23收到修改稿

1 帧间相关性的模型化

在传统的隐马尔可夫模型中(以下仅讨论自左向右无跳转的马尔可夫过程),我们认为,任一时刻出现某观察输出矢量的概率仅依赖于系统当前所处的状态,而与系统在以前时刻所处的状态、观察输出矢量没有关系,其拓扑图如图 1(a)所示. 即系统在 t 时刻出现观察矢量 Y_t 的概率为 $P(Y_t|q_t,\lambda)$,用 $b_{q_t}(Y_t)$ 来刻画.



(a)



(b)

图1

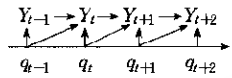


图2

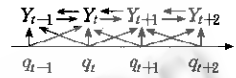


图3

S. Takahashi 等人提出的 BC HMM 认为,任一时刻出现某观察输出矢量的概率不仅依赖于系统当前所处的状态,而且依赖于系统在前一时刻出现的观察矢量,如图 1(b)所示. 即系统在 t 时刻出现观察矢量 Y_t 的概率为 $P(Y_t|Y_{t-1},q_t,\lambda)$,用 $b_{q_tY_{t-1}}(Y_t)$ 来刻画. 为了实现模型参数的可估计化以及参数估计的可靠性,BC HMM 在训练时仅刻画 $P(Y_t|Y_{t-1})$ 和 $b_{q_t}(Y_t)$,进而,根据这两个式子来计算 $b_{q_tY_{t-1}}(Y_t)$,从而在识别时采用 $b_{q_tY_{t-1}}(Y_t)$ 替代训练时使用的 $b_{q_t}(Y_t)$. 由此避免了 Paliwal 提出的模型中完全参数化所带来的问题.

无论如何,从更直观的角度来看如图 2 所示的拓扑图,更能反映帧间的相关性. 即任一时刻出现某观察输出矢量的概率不仅依赖于系统当前所处的状态,而且依赖于系统在前一时刻出现的观察矢量和系统在前一时刻所处的状态. 即系统在 t 时刻出现观察矢量 Y_t 的概率为 $P(Y_t|Y_{t-1},q_{t-1},q_t,\lambda)$,用 $b_{q_{t-1}Y_{t-1}q_t}(Y_t)$(其中 q_{t-1} 为系统在 $t-1$ 时刻所处的状态)来刻画. 如同 Paliwal 提出的模型一样,用有限的训练数据完全参数化该模型几乎是不可能的. 为此,我们需要寻求一种近似的方法来计算 $b_{q_{t-1}Y_{t-1}q_t}(Y_t)$.

更进一步地,如图 3 所示,我们可以采用一阶前后各一帧的帧相关模型. 此时,系统在 t 时刻出现观察矢量 Y_t 的概率为 $P(Y_t|Y_{t-1},Y_{t+1},q_t,q_{t-1},q_{t+1},\lambda)$. 为此,我们需要刻画 $b_{q_tY_{t-1}q_{t-1}Y_{t+1}q_{t+1}}(Y_t)$.

下面,我们来讨论如何采用非线性公式来估计 $P(Y_t|Y_{t-1},q_t,q_{t-1},\lambda)$.

$$\begin{aligned}P(Y_t|Y_{t-1},q_t,q_{t-1},\lambda)&=\frac{P(Y_t,Y_{t-1},q_t,q_{t-1}|\lambda)}{P(Y_{t-1},q_t,q_{t-1}|\lambda)}\\&=\frac{P(Y_t|q_t,\lambda)P(q_t|\lambda)P(Y_{t-1},q_{t-1}|Y_t,q_t,\lambda)}{P(Y_{t-1},q_{t-1}|q_t,\lambda)P(q_t|,\lambda)}\\&=\frac{P(Y_{t-1},q_{t-1}|Y_t,q_t,\lambda)}{P(Y_{t-1},q_{t-1}|q_t,\lambda)}P(Y_t\ q_t,\lambda).\end{aligned}\tag{1}$$

令

$$f(Y_{t-1},Y_t,q_{t-1},q_t,\lambda)=\frac{P(Y_{t-1},q_{t-1}|Y_t,q_t,\lambda)}{P(Y_{t-1},q_{t-1}|q_t,\lambda)},\tag{2}$$

则

$$P(Y_t|Y_{t-1},q_t,q_{t-1},\lambda)=f(Y_{t-1},Y_t,q_{t-1},q_t,\lambda)P(Y_t|q_t,\lambda).\tag{3}$$

进一步地,将式(2)右边项中的分母项 $P(Y_{t-1},q_{t-1}|q_t,\lambda)$ 近似为 $P(Y_{t-1},q_{t-1}|\lambda)$,得

$$f(Y_{t-1},Y_t,q_{t-1},q_t,\lambda)\approx\frac{P(Y_{t-1},q_{t-1}|Y_t,q_t,\lambda)}{P(Y_{t-1},q_{t-1}|\lambda)}.\tag{4}$$

然后,我们用一个非线性的估计公式来计算上式中的右边项,即

$$\frac{P(Y_{t-1},q_{t-1}|Y_t,q_t,\lambda)}{P(Y_{t-1},q_{t-1}|\lambda)}\approx h(b_{q_{t-1}}(Y_{t-1}),b_{q_t}(Y_t)).\tag{5}$$

最后,我们得到了 $P(Y_t|Y_{t-1},q_t,q_{t-1},\lambda)$ 的非线性估计公式:

$$P^*(Y_t|Y_{t-1},q_t,q_{t-1},\lambda)=f(Y_{t-1},Y_t,q_{t-1},q_t,\lambda)P(Y_t|q_t,\lambda)\approx h(b_{q_{t-1}}(Y_{t-1}),(b_{q_t}(Y_t))P(Y_t|q_t,\lambda).\tag{6}$$

2 帧相关 HMM

在本节中,我们使用非线性估计的概念把相邻帧间的相关性糅合到经典的 HMM 中. 为了简化其描述,这里

我们讨论一阶前向的帧相关问题,即如图 2 所示的系统产生当前观察输出矢量的概率仅依赖于前一时刻系统所处的状态和观察输出矢量.

帧相关 HMM(以后我们也记作 FCHMM),$\lambda=(N,\pi,A,B,FC)$ 定义如下:

(1) N,模型中的状态数.

(2) $\pi=\{\pi_i\}$,其中 $\pi_i=P[q_1=i]$,$1\leqslant i\leqslant N$ 是模型初始状态为 i 时的概率,$\sum_{i=1}^{N}\pi_i=1$.

(3) $A=\{a_{ij}\}$,$1\leqslant i,j\leqslant N$,$a_{ij}=P[q_{t+1}=j|q_t=i]$.

(4) $B=b_i(O)$是已知某时刻状态为 i 时出现观察矢量 O 的概率密度函数.

在系统中,我们采用由多个正态分布加权叠加而成的概率密度函数.即

$$b_i(O)=\sum_{m=1}^{M}c_{im}N[O,\mu_{im},U_{im}], \tag{7}$$

M 为正态分布的混合个数.

(5) 帧相关概率分布,$FC=\{f(Y_{t-1},Y_t,q_{t-1},q_t)\}$,其中的 $f(Y_{t-1},Y_t,q_{t-1},q_t)$,我们用非线性估计公式 $h(b_{q_{t-1}}(Y_{t-1}),b_{q_t}(Y_t))$来近似计算.在实验中,我们采用 $h(b_{q_{t-1}}(Y_{t-1}),b_{q_t}(Y_t))=(b_{q_{t-1}}(Y_{t-1})/(b_{q_{t-1}}(Y_{t-1})+b_{q_t}(Y_t))$.

在 FCHMM 中,$P[O_1O_2\ldots O_T|\lambda)]$可由下式求得

$$\begin{aligned}P[O_1O_2\ldots O_T|\lambda)]&=\sum_{\psi=(q_1,q_2,\ldots,q_T)}P[O_1O_2\ldots O_T,\psi|\lambda]=\sum_{\psi}[P(\psi|\lambda)P(O_1O_2\ldots O_T|\psi,\lambda)]\\&=\sum_{\psi}[P(\psi|\lambda)\prod_{t=1}^{T}P^*(O_t|O_{t-1},q_{t-1},q_t,\lambda)]\\&=\sum_{\psi}[P(\psi|\lambda)\prod_{t=1}^{T}f(O_t|O_{t-1},q_{t-1},q_t)b_{q_t}(O_t)]\\&\approx\sum_{\psi}[P(\psi|\lambda)\prod_{t=1}^{T}h(b_{q_{t-1}}(O_{t-1}),b_{q_t}(O_t))b_{q_t}(O_t)].\end{aligned} \tag{8}$$

在使用向前向后公式进行参数重估时,$\alpha_t(j)$修改如下:

$$\alpha_t(j)=\sum_{i=1}^{N}\alpha_{t-1}(i)a_{ij}h(b_i(O_{t-1}),b_j(O_t))b_j(O_t)=\sum_{i=1}^{N}\alpha_{t-1}(i)a_{ij}b_{iO_{t-1}j}(O_t),$$

$$\beta_t(j)=\sum_{i=1}^{N}\beta_{t+1}(i)a_{ji}h(b_j(O_t),b_i(O_{t+1}))b_i(O_{t+1})=\sum_{i=1}^{N}\beta_{t+1}(i)a_{ji}b_{jO_ti}(O_{t+1}).$$

考虑到混合高斯密度可以逼近任一概率分布的原理,我们认为,重估后的 B 矩阵刻画的是考虑帧相关后观察矢量输出的概率密度函数,即 $P(Y_t|Y_{t-1},q_{t-1},\lambda)$.

3　复杂度分析

在本节中,我们对非线性估计帧间相关性的时空复杂度进行分析.显然,该模型利用了混合高斯密度可以逼近任一概率分布的原理,因此,仍用 M 个正态分布的加权和来描述 $P^*(Y_t|Y_{t-1},q_{t-1},q_t,\lambda)$,故该模型未增加任何空间复杂度.

无论是在模型训练还是在识别时,糅合了帧间相关性的 HMM 仅增加了计算 $f(Y_{t-1},Y_t,q_{t-1},q_t)$公式的计算量.在采用非线性估计公式 $h(b_{q_{t-1}}(Y_{t-1}),b_{q_t}(Y_t))$来近似计算 $f(Y_{t-1},Y_t,q_{t-1},q_t)$时,我们仅增加了极其有限的加法和除法运算,故对原来 HMM 的计算复杂度基本上没有影响.

在 Paliwal 提出的模型中需要估计的 B 矩阵参数达 M^2N 个,即为原来 HMM 所需估计 B 矩阵参数的 M 倍.而在 S. Takahashi 等人提出的 BC HMM 在训练时需要额外估计 $P(Y_t|Y_{t-1})$,这样就需要多估计 M^2T 个参数.另外,BC HMM 在识别时需要增加 $P(Y_t|Y_{t-1})$的计算,尤其是需要调整混合高斯密度的权重,从而导致识别时计算复杂度大大增加.

4 实验结果和分析

(1) 语音库描述及 HMM 参数描述

实验中所用的语音库是"863 评测"男声语音库.其中测试集 1 由未训练过说话人的发音数据组成,测试集 2 由训练过说话人的未训练发音数据组成.其余全部语音数据组成训练集.

在实验中,我们选用 5 状态自左向右无跳转的 HMM,每个状态下观察矢量的输出概率由 5 个高斯密度的混合来表征.特征参数采用由线性预测分析得到的 16 维倒谱系数.

(2) Viterbi 解码时非线性估计帧间相关性的效果统计

首先,我们在 Viterbi 解码时使用了帧间相关性非线性估计公式,从而利用帧间相关性对解码路径以及出现此路径且系统输出为待识观察矢量序列的概率值进行调整.识别结果如表 1 所示,从中我们可以看出,HMM+FC−Viterbi 的首选识别率较 HMM(Viterbi)提高了 2 个百分点左右.另外,对各测试集待识音节,我们把不同音节模型的 $\prod_{t=2}^{T} f(Y_t|Y_{t-1},q_{t-1},\lambda)$ 进行了排序统计分析,如图 4 所示.其中,横坐标表示正确音节的非线性连乘积出现在 419 个音节内的排序位置,纵坐标表示不同排序位置的音节占全部统计音节的比例.从图 4 中我们可以看出,帧间相关性的非线性估计公式有助于对待识音节的正确识别.

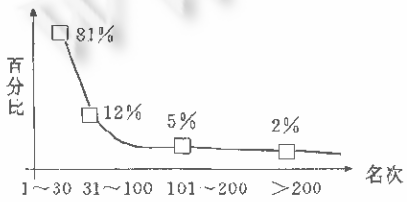


图4 非线性估计公式对待识音节的作用

表 1 识别结果对比表

模 型	识别集	第1名	前2名	前5名	前10名
HMM+(Viterbi)	训练集	59.93	76.29	90.12	95.05
	测试集 1	32.50	48.29	70.05	82.98
	测试集 2	41.22	58.96	80.47	89.88
HMM+(FC−Viterbi)	训练集	62.66	78.42	91.86	95.83
	测试集 1	33.69	48.91	70.08	83.72
	测试集 2	43.90	61.65	82.38	91.13
FCHMM	训练集	66.01	81.10	92.70	96.59
	测试集 1	35.78	51.53	72.02	84.42
	测试集 2	45.63	63.50	82.80	91.23

(3) FC HMM 与 THMM 的识别效果比较

从表 1 中我们可以看出,无论是对训练集还是对测试集,FC HMM 的识别效果均要好于 THMM.对于训练集,FC HMM 首选的识别率比 THMM 要高 6 个百分点;对于测试集 2,即训练说话人的测试集,FCHMM 首选的识别率比 THMM 要高 4 个百分点;对于测试集 1,即未训练过的说话人,FCHMM 首选的识别率比 THMM 要高 3 个百分点.与 S. Takahashi[7]提出的 BC HMM 的识别率相比,FC HMM 中采用非线性估计公式刻画帧间相关性对提高识别率的贡献要稍好于 BC HMM.鉴于 FC HMM 未增加模型的空间复杂度而仅增加了极小的计算量,因此,我们可以说 FC HMM 是一种高效的、在 HMM 中模型化帧间相关性的方法.

5 总 结

在语音识别的声学参数模型化时,声学谱特征和音素的持续特征均需要精确地表达,因为它们都是存在于语音信号中的最基本的特征.本文提出了一种新的方法,用于把时域的相关性糅合到一个基于传统隐马尔可夫模型的语音识别系统中.首先,我们用条件概率的形式处理帧间相关性;然后,用一种非线性概率估计公式来表征相邻帧之间的相关性.目前其他的模型化帧间相关性的方法带来了模型参数和计算量的大幅度增加,与之相比,本文提出的方法丝毫不增加原来 HMM 的空间复杂度,而且也几乎不增加训练和识别阶段的计算复杂度.该方法的另一个特点是,它可以非常方便地糅合到已有的 HMM 之中.值得进一步研究的是,如何更加贴切地非线性估计一阶前向的帧间相关性,以及寻求合适的二阶前后向的帧间相关性估计公式.

参考文献

1 Ostendorf M, Roukos S. A stochastic segment model for phoneme-based continuous speech recognition. IEEE

Transactions on Acoustics, Speech and Signal Processing, 1989,37(12):1857~1869

2 Digalakis V, Rohlicek J R, Ostendorf M. A dynamical system approach to continuous speech recognition. In: Proceedings of the International Conference Acoustics, Speech, and Signal Processing. Mississauga: Imperial Press Limited, 1991. 289~292

3 Wellekens C J. Explicit correlation in hidden Markov model for speech recognition. In: Proceedings of the International Conference Acoustics, Speech, and Signal Processing. San Francisco: IEEE Signal Processing Society, 1987. 383~386

4 Kenny P, Lennig M, Mermelstein P. A linear predictive HMM for vector-valued observations with applications to speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 1990,38(2):220~225

5 Paliwal K K. Use of temporal correlation between successive frames in hidden Markov model based speech recognizer. In: Proceedings of the International Conference Acoustics, Speech, and Signal Processing. San Francisco: IEEE Signal Processing Society, 1993. 215~218

6 Takahashi S. Phonemic HMM constrained by statistical VQ-code transition. In: Proceedings of the International Conference Acoustics, Speech, and Signal Processing. San Francisco: IEEE Signal Processing Society, 1992. 553~556

7 Takahashi S. Phoneme HMM's constrained by frame correlation. In: Proceedings of the International Conference Acoustics, Speech, and Signal Processing. San Francisco: IEEE Signal Processing Society, 1993. 219~222

8 Nam Soo Kim, Chong Kwan Un. Frame-correlated hidden Markov model based on extended logarithmic pool. IEEE Transactions on Speech and Audio Processing, 1997,5(2):149~160

A New Method in Hidden Markov Model for Modeling Frame Correlation

GUO Qing WU Wen-hu FANG Di-tang

(*Speech Laboratory Department of Computer Science and Technology Tsinghua University Beijing* 100084)

Abstract In this paper, the authors present a novel method to incorporate temporal correlation into a speech recognition system based on conventional hidden Markov model (HMM). The temporal correlation is considered to be useful for recognition because of the fact that the speech features of the present frame are highly informative about the feature characteristics of neighboring frames. An obvious way to incorporate temporal correlation is to condition the probability of the current observation on the current state as well as on the previous observation and the previous state. But using this method directly must lead to unreliable parameter estimation for the number of parameters to be estimated may increase too excessively to limited train data. In this paper, the authors approximate the joint conditional PD by non-linear estimation method. As a result, they can still use mixture Gaussian density to represent the joint conditional PD for the principle of any PD can be approximated by mixture Gaussian density. The HMM incorporated temporal correlation by non-linear estimation method, which they called FC (frame correlation) HMM does not need any additional parameters and it only brings a little additional computing quantity. The results of the experiment show that the top 1 recognition rate of FC HMM has been raised by 6 percent compared to the conventional HMM method.

Key words Continuous hidden Markov model (CHMM), frame correlation, non-linear estimation, mixture Gaussian density, joint conditional probability density.