一种新的 HMM 训练方法

贺前华,陆以勤,韦 岗
(华南理工大学电子与通信工程系,广州 510641)

摘 要: 本文是对 HMM 最大距离训练方法的一种改进,该方法采用了更合理的模型距离定义,能更有效地利用训练数据集中的区别信息,使有限的训练数据得到更好的应用,达到提高语音识别系统性能的目的. 导出了 HMM 模型参数的迭代公式. 基于 TIMIT 数据库的非连续语音及连续语音实验结果表明,改进训练方法在降低错识率上较原来的方法有明显改善.

关键词: 隐马尔可夫模型 (HMM); 训练方法; 判决信息
中图分类号: TN911.72 **文献标识码**: A **文章编号**: 0372-2112 (2000) 09-0056-03

A New Approach for HMM Training

HE Qian-hua, LU Yi-qing, WEI Gang
(*Dept. of Electronics Engineering, South China Univ. of Technology, Guangzhou* 510641, *China*)

Abstract: An improved maximum model distance approach was proposed to train HMM. By adopting a more realistic model distance definition, discriminative information contained in the training data could be used to improve the performance of recognizer. HMM parameter adjustment rules were induced. Both isolated word and continuous speech recognition experiments on TIMIT database showed that significant error reduction could be achieved by the improved approach.

Key words: Hidden Markov Model (HMM); training approach; discriminative information

1 引言

隐马尔可夫模型(HMM)在语音识别,特别是在连续语音识别领域得到了广泛应用[1,3,4],其根本问题是如何利用有限的训练数据建造一个最佳的识别器,即模型训练问题. 除经典的最大似然准则(ML)外,已提出了一系列的训练方法,如最大互信息法[5]、最小鉴别信息法[6]、最小识别误差法[7]等. 这些方法都有其自身的优越性及局限性;引文[2]则从利用语音帧间的相关信息来提高基于 HMM 的语音识别系统的性能. 文[1]提出了一种基于最大模型距离的 HMM 训练方法(MMD),该方法原理上可利用不同类训练数据中的区别信息,使系统的性能得到提升. 设训练任务为利用数据 $O=\{O^1, O^2, \cdots, O^M\}$ 建立识别器 $\Lambda=\{\lambda_v, v=1, \cdots, M\}$,其中 O^v 为模型 λ_v 的标记训练数据,定义 $\bar{O}^v = O - O^v$ 为 λ_v 的竞争训练数据,MMD 将所有的竞争训练数据一视同仁. 事实上,不同竞争训练数据对 λ_v 的竞争程度是不同的,应该区别对待. 为了解决这一问题,本文提出了一个更合理的模型距离定义,并称这一定义下的 MMD 算法为改进 MMD (IMMD). 实验结果表明,IMMD 比 MMD 有更佳的性能表现.

2 改进的最大模型距离方法(IMMD)

先简单回顾一下 MMD 方法的基本思想[1],MMD 首先将模型对 λ_v 及 λ_θ 的概率距离 $D(\lambda_v, \lambda_\theta)$[8] 推广到有限长观察序列:

$$D(\lambda_v, \lambda_\theta) = \frac{1}{T_v}\{\log P(O^v|\lambda_v) - \log P(O^v|\lambda_\theta)\} \quad (1)$$

其中 $O^v = (o_1^v, o_2^v, o_3^v, \cdots, o_{T_v}^v)$. 并在此基础上定义模型 λ_v 与模型集 Λ 的距离度量 $D(\lambda_v, \Lambda)$ 为:

$$D(\lambda_v, \Lambda) = \frac{1}{T_v}\left\{\log P(O^v|\lambda_v) - \frac{1}{V-1}\sum_{\theta=1,\theta\neq v}^{V}\log P(O^v|\lambda_\theta)\right\} \quad (2)$$

最大模型距离方法按下式优化模型参数.

$$(\Lambda)_{MMD} = \arg\max_{\Lambda}\sum_{v=1}^{V} D(\lambda_v, \Lambda) \quad (3)$$

式(2)将 λ_v 的所有竞争数据 O^θ 置于同等重要位置,这与下述识别判决规则不一致.

$$C(O) = C_v \text{ iff } P(O|\lambda_v) = \max_\theta P(O|\lambda_\theta) \quad (4)$$

基于判决规则(4),λ_v 的竞争者可分为两类,一类为 $S_1=\{\lambda_\theta, P(O|\lambda_\theta) \geq P(O|\lambda_v)\}$,另一类为 $S_2=\{\lambda_\theta, P(O|\lambda_\theta) < P(O|\lambda_v)\}$. 当 S_1 非空时,识别时就会做出错误判决,模型训练的目的就是尽量使 S_1 为空. 为此目的,一种直观且合理的方法是在训练进程中使 $P(O^v|\lambda_\theta)$ $(\lambda_\theta \in S_1)$ 下降. 结合判决训练的基本思想,我们重定义 $D(\lambda_v, \Lambda)$ 如下:

收稿日期:1999-04-21;修回日期:2000-04-08
基金项目:国家自然科学基金(No. 69881001);广东省自然科学基金(No. 970445)

$$D(\lambda_v,\Lambda)=\frac{1}{T_v}\left\{\log P(O^v|\lambda_v)-\log\left[\frac{1}{M-1}\sum_{\theta=1,\theta\neq v}^{M}(P(O^v|\lambda_\theta))^{\eta}\right]^{1/\eta}\right\} \tag{5}$$

其中 η 为一正加权系数. 通过改变 η 的大小,λ_v 的竞争者对 $D(\lambda_v,\Lambda)$ 的影响程度会有不同的分配. 当 η 愈小时,各竞争者对 $D(\lambda_v,\Lambda)$ 的影响愈趋一致,而当 η 愈大时,S_1 中的竞争者对 $D(\lambda_v,\Lambda)$ 的影响愈大,S_2 中的单元对 $D(\lambda_v,\Lambda)$ 的影响愈小. 特别是当 η 趋近于 ∞ 时,上式 $[\cdot]^{1/\eta}$ 一项变为 $\max_{\theta=1,\theta\neq v}P(O^v|\lambda_\theta)$,此时只有最大竞争者被考虑. 定义 $D(\Lambda)=\sum_{v=1}^{M}D(\lambda_v,\Lambda)$,IMMD 的基本思想是寻找使 $D(\Lambda)$ 最大的模型参数. $D(\Lambda)$ 的极大值可用梯度法求得,参数迭代公式为:

$$\Lambda_{n+1}=\Lambda_n+\varepsilon_n U_n\nabla D(\Lambda)\big|_{\Lambda=\Lambda_n} \tag{6}$$

这里用 Λ 区别于满足概率约束条件的 Λ. ε_n 为满足随机收敛条件[9]的正数. U_n 为单位矩阵或适当选取的正定矩阵. 对于 N 状态的连续 HMM,其状态输出分布概率为:

$$b_j(o)=\sum_{k=1}^{K}c_{jk}N(o,u_{jk},R_{jk}),\ 1\le j\le N,$$

其中 c_{jk} 为小于 1.0 的正数,满足条件 $\sum_{k=1}^{K}c_{jk}=1$. $N(\cdot)$ 为高斯分布,平均矢量 $u_{jk}=[u_{jkl}]_{l=1}^{L}$,方差矩阵 R_{jk} 强制为对角矩阵,即 $R_{jk}=[\sigma_{jkl}^2]_{l=1}^{L}$. 由式(6)可导出模型 λ_v 的参数迭代公式为:

$$\tilde{\pi}_i^{n+1}=\pi_i^n+\varepsilon_n\left(\gamma_1^v(i)-\sum_{\theta=1,\theta\neq v}^{M}R_\theta^v\gamma_1^\theta(i)\right),\ i=1,2,\cdots,N \tag{7a}$$

$$\tilde{a}_{ij}^{n+1}=a_{ij}^n+\varepsilon_n\left(s_{ij}^v-\sum_{\theta=1,\theta\neq v}^{M}R_\theta^v s_{ij}^\theta\right),\ i,j=1,2,\cdots,N \tag{7b}$$

$$c_{jk}^{n+1}=c_{jk}^n+\varepsilon_n\left(\gamma^v(j,k)-\sum_{\theta=1,\theta\neq v}^{M}R_\theta^v\gamma^\theta(j,k)\right) \tag{7c}$$

$$u_{jkl}^{n+1}=u_{jkl}^n+\varepsilon_n\left(\Delta o_l^v-\sum_{\theta=1,\theta\neq v}^{M}R_\theta^v\Delta o_l^\theta\right) \tag{7d}$$

$$\sigma_{jkl}^{n+1}=\sigma_{jkl}^n+\varepsilon_n\sigma_{jkl}^n\left(\phi_{jkl}^v-\sum_{\theta=1,\theta\neq v}^{M}R_\theta^v\phi_{jkl}^\theta\right) \tag{7e}$$

其中:$R_k^v=P^\eta(O^k|\lambda_v)\Big/\left(\sum_{\theta=1,\theta\neq k}^{M}P^\eta(O^k|\lambda_\theta)\right)$,

$$\gamma_1^\theta(i)=P(q_1=i|O^\theta,\lambda_v)/T_\theta,$$

$$s_{ij}^\theta=\frac{1}{T_\theta}\sum_{t=1}^{T_\theta-1}P(q_t=i,q_{t+1}=j|O^\theta,\lambda_v),$$

$$\gamma^\theta(j,k)=\frac{1}{T_\theta}\sum_{t=1}^{T_\theta}\gamma_t^\theta(j,k),$$

$$\gamma_t(j,k)=\left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)}\right]\left[\frac{c_{jk}N(o_t,u_{jk},R_{jk})}{\sum_{k=1}^{K}c_{jk}N(o_t,u_{jk},R_{jk})}\right]$$

$$=\begin{cases}\dfrac{1}{P(O|\lambda)}\pi_j\beta_1(j)\,c_{jk}N(o_1,u_{jk},R_{jk}), & t=1\\ \dfrac{1}{P(O|\lambda)}\sum_{i=1}^{N}\alpha_{t-1}(i)\,a_{ij}\beta_t(j)\,c_{jk}N(o_i,u_{jk},R_{jk}), & t>1\end{cases}$$

$$\Delta o_t^\theta=\frac{1}{T_\theta}\sum_{t=1}^{T_v}\gamma_t^\theta(j,k)(o_{tl}^\theta-u_{jkl}),$$

$$\phi_{jkl}^\theta=\frac{1}{T_\theta}\sum_{t=1}^{T_\theta}\gamma_t^\theta(j,k)\left[\left(\frac{o_{tl}^\theta-u_{jkl}}{\sigma_{jkl}}\right)^2-1\right]$$

式(7)与 MMD 的模型参数重估公式[1]的差别在于加权系数 R_k^v 的引入,而 R_k^v 恰恰反映了 O^k 对 λ_v 参数重估的影响份量. 设 $P_{vk}=P(O^k|\lambda_v)$,则 $R_k^v=\left(\frac{P_{vk}}{P_{kk}}\right)^{\eta}\Big/\sum_{\theta=1,\theta\neq k}^{M}\left(\frac{P_{\theta k}}{P_{kk}}\right)^{\eta}$. 由于 P_{vk}/P_{kk} 是模型 λ_v 与 λ_k 相似度的一种度量,因此 R_k^v 可解释为 λ_v 与 λ_k 之间的一种相对相似性度量,R_k^v 大,表明 λ_k 与 λ_v 的相似性比 λ_k 与其它竞争者的相似性要大,反之亦然. 当 λ_v 与 λ_k 的其它竞争者相比更象 λ_k 时,O^k 与 O^v 之间的区别信息得到更好的利用,从而提高 λ_v 辨识 O^v 与 O^k 的能力.

当 η 逼近 ∞ 时,若 O^k 使 λ_v 成为 λ_k 的最强竞争者,$R_k^v=1$,否则,$R_k^v=0$,当 η 逼近 0 时,R_k^v 变为 $1/(M-1)$,此时 IMMD 退化为 MMD.

3　实验及结果

我们孤立词识别及连续语音识别两方面来评估 IMMD 的性能.

3.1　孤立词识别

词汇集包括 21 个英文单词,这些词是 TIMIT 语音数据库中 SA1 及 SA2 语句的组成词. 每个词有 244 个语音样本,其中 160 个做为 HMM 模型训练,另外 84 个做为测试用. HMM 为 6 状态连续 left-to-right 模型,每个状态的输出概率分布由 3 个高斯密度混合组成. 表 1 给出了分别用 IMMD、MMD、ML 方法训练的识别器的识别错误率. 该结果表明 IMMD 比 MMD 好,MMD 比 ML 好. 相对 MMD 而言,IMMD 的开集错识率下降了 17.96 %,闭集错识率下降了 51.44 %.

表 1　孤立词识别实验结果(错识率)

	ML	MMD	IMMD	错识下降比率
闭集测试	2.14 %	1.73 %	0.84 %	51.44 %
开集测试	2.43 %	2.06 %	1.69 %	17.96 %

3.2　连续语音识别

连续语音识别实验为识别 TIMIT 语音数据库的 60 个音素(不包括静音). 由于 TIMIT 的语音除两端外,不包含静音成份,因此实验中不考虑静音,而语音的起点及终点 TIMIT 已有详细标注. 从 TIMIT 中选取了 600 句构成一音素数据量较平衡的测试数据库(音素频次 23147),其中 400 句作为训练数据(音素频次 15688),另 200 句作为集外测试数据. 每个音素建立一个与上下文无关的 3 状态 HMM 模型,每个状态混合密度数为 5. 训练进程包括两个阶段:第一阶段利用已切分的数据建立各音素的声学统计模型;第二阶段利用已训练的音素模型对语句进行强制识别切分,从而得到新的切分数据. 上述两阶段交替进行,直到满足预设的收敛条件. 最初的语句切分为简单均一切分法.

设 Cor 及 Del,Ins,Sub 分别 W 表示正确识别的音素个数,删除、插入、替换次数及测试数据中的总音素量,其中 $Cor=W-Del-Sub$. 定义正识率为 $\%Corr=(Cor/W)\times100\%$,准确率为 $\%Acc=[(Cor-Ins)\times100\%/W]$. 表 2 给出了 IMMD、MMD 及 ML 方法的识别器性能. 再一次证明了 IMMD 比 MMD 好.

表 2 连续语音识别实验结果

训练方法	1 - %Corr	1 - %Acc
ML	14.81 %	19.78 %
MMD	14.02 %	18.95 %
IMMD	13.15 %	18.04 %
错识率下降比率	6.21 %	4.80 %

4 结论

本文讨论了 MMD 方法[1]的缺点,结合判决训练方法的基本思想,提出了更合理的模型距离定义,导出了 CHMM 的参数迭代公式,从而实现了非标记数据 O^{θ} 对其竞争模型参数重估的影响强度与竞争模型对正确识别出 O^{θ} 的影响程度相关联,这一关系的建立,使 IMMD 可比 MMD 更有效地利用训练数据中隐含的判决信息提高识别器的总体性能. 基于 TIMIT 的孤立词识别及连续语音识别实验均证实了这一推断. 在孤立词识别实验中,IMMD 将 MMD 的闭集及开集错误率降低了 51.44 %和 17.96 %,在连续语音识别实验中,MMD 的开集正识率及准确率分别改善了 6.21 %和 4.80 %.

MMD 是 IMMD 在 $\eta = 0$ 时的特殊情形. IMMD 方法除应用于 HMM 训练外,其基本思想还可应用于更一般的统计模型的训练问题,如分段模型[4]和随机轨迹模型[10].

参考文献:

[1] 贺前华,韦岗. 基于模型距离的 HMM 训练方法 [J]. 计算机工程,1996,22(6):471 - 476.

[2] 杨浩荣,王作英,陆大[illegible]. 语音识别 HMM 中引入帧间相关信息的一种参数化模型 [J]. 电子学报,1998,26(10):50 - 54.

[3] Y. Gotoh, M. M. Hochberg, H. F. Silverman. Efficient training algorithms for HMM's using incremental estimation [J]. IEEE Trans. on Speech and Audio processing. November 1998,6(6):539 - 548.

[4] M. Ostendorf, V. V. Digalakis, O. A. Kimball. From HMM's to segment models: a unified view of stochastic modeling for speech recognition [J]. IEEE Trans. on Speech and Audio processing. 1996,4(5):360 - 378.

[5] R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer. Maximum mutual information estimation of hidden Markov model parameters for speech recognition [A]. in Proc. 1986 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing [C], Tokyo, Japan, Apr, 1986:49 - 52.

[6] Yariv Ephraim, Amir Dembo, L. R. Rabiner. A minimum discrimination information approach for hidden Markov modeling [J]. IEEE Trans. on Information Theory, Sept. 1989,35(5):1001 - 1003.

[7] Biing-Hwang Juang, Wu Chou, Chin-Hui Lee. Minimum classification error rate methods for speech recognition [J]. IEEE Trans. on Speech and Audio processing. May. 1997,5(3):257 - 265.

[8] H. Juang, L. R. Rabiner. A probabilistic distance measure for hidden Markov models [J]. AT&T Technical Journal, February 1985,64(2):391 - 408.

[9] P. C. Chang, B. H. Juang. Discriminative template training for dynamic programming speech recognition [A]. Proc. San Francisco, 1992, ICASSP-92, I:493 - 496.

[10] F. Gong. Stochastic trajectory modeling and sentence searching for continuous speech recognition [J]. IEEE Trans. on Speech and Audio processing. 1997,5(1):33 - 44.

作者简介:

贺前华 1965 年出生,1987 年于湖南师范大学本科毕业,1990 年于西安交通大学获硕士学位,1993 年于华南理工大学获博士学位并留校工作,1994 年至 1995 年在香港城市大学工作,现为华南理工大学副教授. 从事语音及音频信号处理、进化算法及自然语言处理方面的研究.

陆以勤 1968 年生,1994 年至 1995 年到香港城市大学联合培养,1996 年于华南理工大学获博士学位并留校工作,从事语音识别、Petri 网及其应用研究.