基于经典隐马尔可夫模型的汉语连续语音识别系统 [1]

郝 杰 李 星

(清华大学电子工程系 100084)

摘 要 该文构造了基于经典隐马尔可夫模型 (Hidden Markov Model, HMM) 的汉语连续语音识别系统，定量地分析与评价了经典 HMM 的性能.

关键词 汉语连续语音识别，经典 HMM

中图号 TP391.42, TN912.3

1 引 言

目前，国内对隐马尔可夫模型 (HMM) 的理论研究已经具备了国际先进水平，并且成功应用于汉语连续语音识别. 1998 年 4 月，在第五届 863 评测 [1] 中，清华大学电子工程系开发的连续语音识别系统 THEESP(TsingHua Electronic Engineering SPeech) 的汉字正确率达到了 98.7%，拼音首选识别率达到 79.4%. 该系统采用的基于段长分布的 HMM(Duration Distribution Based-HMM， DDBHMM) 从理论上解决了 HMM 的状态驻留长度的建模问题 [2]，并且解决了非齐次的 DDBHMM 的搜索问题 [3]，是对经典 HMM 的推广.

另一方面，定量地评价经典 HMM 应用于汉语连续语音识别时的性能，以及 DDBHMM 相对于经典 HMM 在识别性能上的提高，是汉语语音识别这一领域亟需解决的问题.

2 经典 HMM 的实现

本文实现了经典 HMM 的 Baum-Welch 训练与 Viterbi 搜索，利用 98'863 测试提供的语音数据库完成识别实验. 构造识别系统时，考虑到汉语普通话的语言学和语音学特点 [4]，同时为了评价经典 HMM 在声学层的性能，采用静态识别网络研究声学层识别，即语音到拼音的识别. 为了考察语音单元的选择及 HMM 模型的参数个数对识别性能的影响，本文训练并测试了 3 种半音节单元集合：第 1 种由 21 个 Initial 和 39 个 Final 组成，也就是“汉语拼音方案”中的所有声母和韵母；第 2 种由 95 个 Initial 和 39 个 Final 组成，Initial 的选取考虑了声母与后接韵头的搭配；第 3 种由 94 个 Initial 和 170 个 Final 组成，Final 是带调 (四声) 的韵母. 为了研究输出概率密度函数的形式及 HMM 的参数个数对识别性能的影响，本文训练并测试了 3 种输出概率模型：第 1 种是协方差矩阵为对角阵的单高斯分布；第 2 种是协方差为满阵的单高斯分布；第 3 种是多个协方差矩阵为对角阵的单高斯分布的线性组合，即混合高斯分布. 本文后续各部分通过一些实验，针对语音单元和输出概率的各种组合，研究经典 HMM 的复杂度、可靠性、精确性与训练集合的数据量、训练时间、解码效率等特性之间的关系；并且通过实验分析多候选的构造和剪枝的意义.

下述所有实验都采用美尔 (Mel) 刻度的倒谱系数 (Melscale Frequency Cepstrum Coefficient， MFCC)[5] 作为语音特征，8 个数据集合均选自 863 语音数据库，见表 1.

3 非特定人连续语音识别实验

采用 SI-TRN1 作为训练集，SI-TST1 作为测试集，识别结果见表 2，识别率是在最优的惩罚概率 (对数值) 下得到的 (下同). 本文试图比较具有满协方差阵的单高斯分布和 12 个具有对角协方差阵的单高斯的线性组合的性能，由于这两种分布的参数个数分别是 1081 与 1104，比较接近，所以在模型参数个数相当的前提下，认为这两种分布下的识别率具有可比性. 在理

[1] 2000-09-25 收到，2001-07-11 定稿

国家自然科学基金，国家杰出青年科学基金 (69625103) 资助项目

论上，HMM 状态在特征空间中的统计分布不一定是单峰的，所以 12 个混合高斯分布应该比满协方差阵单高斯分布对状态输出概率描述得更准确，如果训练数据充足，前者的识别率应该一致地高于后者.

表 1　数据集合

数据集	用途	人数	句数 (音节数)	数据长度
SD-TRN	特定人训练集	1	479	41.58min
SD-TST	特定人测试集	1	40(511)	3.59min
SI-TRN1	非特定人训练集	35	18718	24.06h
SI-TST1	非特定人测试集	6	120(1536)	8.32min
SI-TRN2	非特定人训练集	69	40624	47.12h
SI-TST2	非特定人测试集	14	280(3004)	17.08min
SI-TST3	非特定人测试集	6	240(3145)	17.15min
SI-TST4	非特定人测试集	3	60(465)	3.65min

表 2　非特定人连续语音识别结果

输出概率密度形式	识别率 (最优惩罚概率的对数值)		
	语音单元数 =60	语音单元数 =134	语音单元数 =264
对角协方差阵单高斯分布	40.76% (-40)	52.15% (-50)	56.97% (-50)
2 个混合高斯分布	45.90% (-40)	57.68% (-50)	62.17% (-50)
4 个混合高斯分布	53.19% (-50)	64.45% (-40)	68.03% (-40)
8 个混合高斯分布	58.85% (-50)	69.34% (-40)	72.85% (-40)
满协方差阵单高斯分布	58.66% (-40)	68.83% (-40)	73.76% (-40)
12 个混合高斯分布	61.13% (-50)	70.83% (-40)	74.09% (-50)
16 个混合高斯分布	62.17% (-50)	72.07% (-50)	75.65% (-40)
24 个混合高斯分布	------	------	76.50% (-50)
32 个混合高斯分布	------	------	76.63% (-50)

由表 2 可见，对于语音单元数和输出概率密度形式的各种组合，训练集 SI-TRN1 的数据量都是充足的，随着语音单元数即 HMM 个数的增加，和 (或) 状态输出概率模型的参数个数的增加，识别率一致地提高；12 个混合高斯分布的识别率一致地高于满协方差阵单高斯分布的识别率，而且 8 个混合高斯分布的识别率就已经逼近甚至超过了满协方差阵单高斯的识别率，但是后者的训练时间和识别时间 (都是在 Pentium II 400MHz 的计算能力下测得，下同) 都一致地少于前者，见表 3，说明了在计算复杂度上，满协方差阵单高斯分布优于混合高斯分布；各种组合下，最优惩罚概率的取值 (对数值) 稳定在了 $[-50,-40]$ 之间.

表 3　非特定人连续语音训练时间与识别时间

输出概率密度形式	一次 Baum-Welch 迭代时间 /Viterbi 搜索时间 (min)		
	语音单元数 =60	语音单元数 =134	语音单元数 =264
8 个混合高斯分布	509/10	510/15	515/39
满协方差阵单高斯分布	370/9	397/14	427/38

为了考察训练数据量对识别率的影响，本文利用数据集 SI-TRN2 训练了单元数为 264 的满协方差阵单高斯分布的 HMM，记作 FULLC2，表 2 中对应组合下的 HMM 记作 FULLC1. 利用表 2 中语音单元数为 264 的 9 种 HMM 及 FULLC2，共 10 种 HMM，在测试集 SI-TST1 确定的最优惩罚概率 (见表 2 第 3 列) 下，对本文中的 4 个非特定人测试集进行识别实验，结果见图 1. 由图 1 可以看出，测试集 SI-TST3 的识别率一致地高于其它测试集，原因是这部分语音数据是 98'863 测试用于听写机测试的，其录音质量非常高；测试集 SI-TST4 的识别率一致地低于其它测试集，原因是这部分语音数据是 98'863 测试用于声学层测试的，是由无意义的音节组合构成的短句，表 2 中的最优惩罚概率与这个测试集失配，插入错误显著；SI-TST1，SI-TST2 两个测试集的识别率介于以上两者之间. 由此可见，语音质量对识别率有重要的影响. 为了考察说话人对识别率的影响，测试集 SI-TST2 中的 7 个说话人的识别率见图 2，最高识别率 (M46) 比最低识别率 (M96) 高出 25%~30%，由此可见说话人的差异对识别率具有更大的影

响；测试集 SI-TST1， SI-TST2 中的说话人是从 863 数据库中随机抽选的，图 2 中识别率依说话人的分布也说明了图 1 中对应的平均识别率是稳定、可靠的.

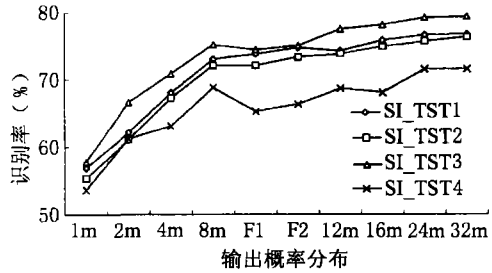

图 1 4 个测试集对 10 种 HMM 的识别率
(m 为 mix, F 为 FullC, 以下各图相同)

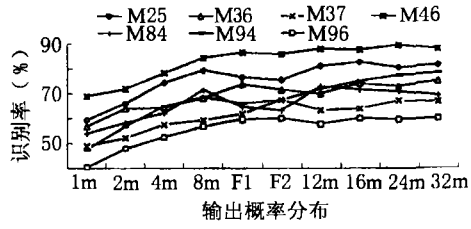

图 2 测试集 SI-TST2 中 7 个说话人的识别率

4 拼音多候选的构造

利用图 1 中的 10 种 HMM，可以作拼音多候选的 Viterbi 搜索，多候选的构造基于 Lattice-Dependent 算法 [6]，即在 Viterbi 搜索完成后，在首选的时间切割点直接回溯出多候选，各候选依该时刻的积累概率排序. 图 3 给出测试集 SI-TST3 的前 10 选识别率，最优惩罚概率是由测试集 SI-TST1 确定的 (见表 2 第 3 列)．可以看出多候选对识别率有显著的提高，由于声学层识别的搜索空间比较简单， 10 候选即可达到 90% 以上的识别率.

98'863 测试的声学层测试是在已知句长的前提下对测试集 SI-TST4 的多候选识别，图 4 给出了已知句长时，图 3 中 10 种 HMM 对 SI-TST4 的 5 候选识别率，以及 98'863 测试的冠军系统 THEESP[1] 的识别率. 由于已知句长排除了插入与删除错误的发生，所以图 4 中 10 种 HMM 的识别率一致地高于图 1 中的结果；另一方面， 10 种 HMM 的识别率都低于 THEESP，原因主要是 THEESP 采用了更多的语音单元 (三音子， triphone)[7] 以及更多的训练数据.

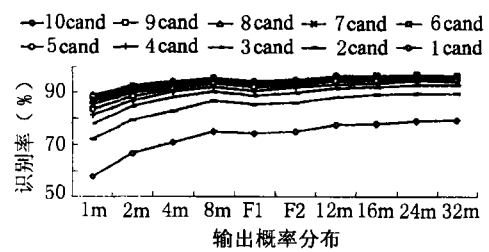

图 3 测试集 SI-TST3 的前 10 候选识别率

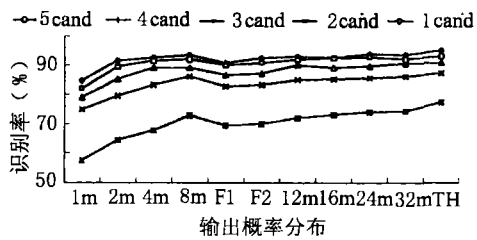

图 4 已知句长时测试集 SI-TST2 的 5 个候选识别率
(TH 为 THEESP)

5 剪 枝

为了提高识别速度，通常要在搜索过程中进行剪枝处理. 图 5，图 6 给出了对测试集 SI-TST2 进行束搜索剪枝的结果，即束搜索宽度分别为 200.0， 150.0， 100.0 时的搜索时间与首选识别率. 在束搜索宽度分别为 100.0 时，搜索速度提高约一倍，代价是识别率下降 3 到 5 个百分点. 由于在不显著降低识别率的前提下，剪枝不能显著提高搜索速度，所以声学层识别中剪枝没有太大的意义. 在声学层与语言模型两者合一的语音识别系统中，搜索空间很复杂，各种剪枝算法能够有效地限制搜索空间，提高搜索速度.

6 结 论

本文通过以上的汉语连续语音识别实验及分析，可以对经典 HMM 得出以下结论：

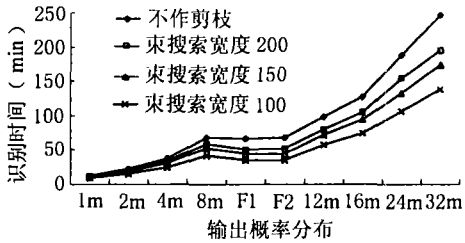

图 5　不同剪枝门限下 SI-TST2 的识别时间

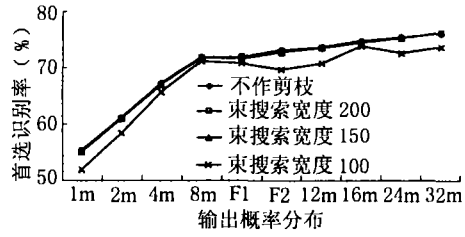

图 6　不同剪枝门限下 SI-TST2 的首选拼音识别率

(1) 在训练数据充足的前提下，语音单元数和输出概率模型参数的增加及语音质量的提高对识别率有一致的提高；

(2) 在输出概率模型参数的个数相当的前提下，混合高斯分布比满协方差阵的单高斯分布的描述能力强，识别率高；但是在训练与识别速度上，后者具有优势；

(3) 汉语语音的声学层识别中拼音多候选的构造是一个容易解决的问题；

(4) 汉语语音的声学层识别对应的搜索空间的复杂度较低，剪枝的意义不大.

本文所构造的系统与具有国内领先水平的 THEESP 系统的识别率相当，所得实验结果和结论为汉语语音识别的深入研究提供了必要的参考和依据. 尚需解决的任务是：针对经典 HMM 的齐次假设的不合理性，考察经典 HMM 与非齐次 DDBHMM 之间的性能差异.

参　考　文　献

[1] 863 办公室，第五届全国汉字识别、语音识别与合成系统及自然语言处理系统评测结果，智能机研究动态，1998, 4.

[2] 王作英，基于段长分布的 HMM 语音识别模型，第二届全国汉字语音识别会议，庐山，1989.

[3] Wang Zuoying, Gao Hongge, An inhomogeneous HMM speech recognition algorithm, Proceedings of the first international conference on multimodal interface(ICMI'96), 1996, Beijing, 70–74.

[4] Lee Linshan, Voice dictation of mandarin Chinese, compuer data entry without a keyboard via speech recognition. IEEE Signal Processing Magazine, 1997, 7, 63–101.

[5] J. W. Picone, Signal modeling techniques in speech recognition, Proc. of the IEEE, 1993, 81(9), 1215–1247.

[6] R. Schwartz, S. Austin, A comparison of severel approximate algorithms for finding multiple (N-Best) sentence hypotheses, Proceedings of ICASSP, 1991, Toronto, 701–704.

[7] 赵庆卫，非特定人大词汇量汉语连续语音识别系统的研究，[博士学位论文]，北京，清华大学电子工程系，1998.

MANDARIN CONTINUOUS SPEECH RECOGNIZER BASED ON CLASSICAL HMM

Hao Jie　　Li Xing

(*Dept. of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

Abstract　In this paper a mandarin continuous speech recognition system is implemented based on classical Hidden Markov Model (HMM). The performance of classical HMM is quantitatively analyzed and evaluated.

Key words　Mandarin continuous speech recognition, Classical HMM

郝　杰：　男，1973 年生，博士生，从事语音识别研究.

李　星：　男，1956 年生，教授，博士生导师，从事信号处理、网络、多媒体通信等研究与教学.