汉语语音听写机技术的研究与实现*

郑 方 牟晓隆 徐明星 武 健 宋战江

(清华大学计算机科学与技术系语音实验室 北京 100084)

E-mail: fzheng@sp.cs.tsinghua.edu.cn

摘要 文章从声学基元和词法树两个方面对连续语音识别和汉语语音听写机中声学层面的搜索策略进行了分析,提出了基于统计知识的帧同步搜索算法和基于词法约束的词搜索树结构,构成了声学层面的双层搜索网络.算法中利用了统计知识,包括声学层面的差分状态驻留信息和特征变化量信息等.实验结果表明,基于知识的搜索策略使连续语音识别的性能提高了36.6%.文章还介绍了*N*-Gram统计语言模型的修正退化频度估计算法和搜索算法原理.通过对多年研究成果的分析,实现了一个汉语语音听写机的引擎,并在PC机上构建了两个系统:非特定人汉语语音听写机实用编辑器ST97和语音命令系统CMD97.

关键词 连续语音识别,汉语语音听写机,搜索策略,基于统计知识的帧同步搜索算法,差分状态驻留,特征变化量,词搜索树,双层搜索网络,汉语语音听写机引擎,语音命令.

中图法分类号 TP391

在人机交互技术中,人机语音对话是一种最为自然的方式.在这样的系统里,人们通过语音向机器发出指令或利用语音录入文本,机器则根据识别结果执行用户发出的命令或保存语音录入的文本,并用语音给出响应.在这项技术中,连续语音识别是一个非常重要的环节.

首先,我们分析和研究了汉语语音听写机中所采用的核心技术,包括特征提取、声学模型及基元选择,搜索算法和语言模型等,并针对其中的难点进行了分析和实验,提出了新的高效算法,包括基于统计知识的帧同步搜索(SKB-FSS)算法、受词法约束的词搜索树(WST)、连续语音识别中的双层搜索网络(TLSN)、*N*-Gram的修正退化频度估计算法等.其次,我们设计并实现了在PC机上可以运行的汉语语音听写机引擎(CDME).利用该引擎,可以方便地进行语音听写系统和语音命令系统的应用开发.最后,文章对目前汉语语音听写机研究中的不足进行了分析,指出了作者以后的研究方向和重点.

1 汉语语音听写机的核心技术

作为实现汉语语音听写机(Chinese dictation machine,简称CDM)的基础,本节对CDM中的核心技术进行详细的研究和讨论,并提出了一些新的、切实可行的方法.

1.1 特征提取、声学模型及基元选择

在16KHz采样频率下,语音的特征选为16维的倒谱系数及其线性回归分析系数(简称自回归系数)[1,2],线性回归分析宽度为5帧.基于以往的实验,我们假定语音的倒谱系数和自回归系数是相互独立的,对两者分别建模.

---

* 本文研究得到国家863高科技项目基金资助.作者郑方,1967年生,博士,副教授,主要研究领域为信号处理,语音识别与合成,语言理解,关键词识别.牟晓隆,1973年生,硕士,主要研究领域为语音识别.徐明星,1973年生,博士生,主要研究领域为语音识别.武健,1975年生,硕士生,主要研究领域为语音识别.宋战江,1972年生,博士生,主要研究领域为语音识别和信号处理.

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

本文1998-02-24收到原稿,1998-05-12收到修改稿

对传统的 HMM 来说,用以对特征空间进行描述的 B 矩阵是非常重要的,它所扮演的角色要比状态转移矩阵 A 重要得多[3~7]. 特征空间的描述通常采用混合高斯密度(mixed Gaussian densities,简称 MGD)[8,9]或栓柱式混合高斯密度(tied MGD)[9,10]. 著名的 Viterbi 解码算法[11]和帧同步算法[12]则分别提供任意拓扑结构的 HMM 和从左向右结构的 HMM 的状态解码方案(该过程称为搜索).

作者通过大量的实验,采用修改的 HMM 模型——中心距离连续概率模型(center distance continuous probability model,简称 CDCPM)[13,14]——对语音进行建模. CDCPM 是去掉了 A 矩阵的从左向右 HMM. 在没有 A 矩阵的情况下,状态转移的控制有所不同. 对训练,由高效的、高鲁棒性的非线性分块(non-linear partition,简称 NLP)算法[15,16]提供状态序列;对识别,则采用修改的 Viterbi 解码算法[16,17]. 在状态空间的刻画方面,以高鲁棒、低存储的中心距离正态(center distance normal,简称 CDN)分布替代高斯正态分布,该分布对正态随机向量离开其中心的距离进行刻画. 尝试了采用基于最近邻(NN)原则的嵌入式多模板(embedded multiple-model,简称 EMM)方案描述和混合 CDN 密度描述[13,14],并对密度数目的选取进行了研究[18],压缩了存储,提高了性能,取得了较好的效果.

在基元的选取方面,汉语语音有 400 多个无调音节和四声,共计 1 300 多个有调音节[19]. 每个音节由声母和韵母组成,其中声母(包括零声母)22 个,韵母 38 个. 汉语的音素可以更细地分为 19 个元音音素和 22 个辅音音素. 有专门的实验比较了这 3 种识别单元,其结论是:识别单元选取得越小,模型时空开销越小,灵活性越大,但训练数据库的标定就越困难,识别性能也越差;而识别单元选得越大,灵活性就越小,但识别效果越好. 两者之间存在一个折衷的选择,但以音节为佳,可同时满足性能和灵活性的要求[17].

1.2　基于统计知识的帧同步搜索(SKB-FSS)算法

在声学层面,每个基元对应一个模型,通常是传统 HMM 或修改的 HMM,它们大多采用从左向右的拓扑结构. 在这种结构下使用的帧同步算法[12]或修改的帧同步算法[16](包括 Viterbi 解码算法)都是基于动态规划最优原理的,即"时刻 t 达到结点 n 的最佳路径,可以由时刻 $t-1$ 到达任一结点 m 的最佳路径和 t 时刻的最佳策略(结点 m 到结点 n 的转移)来决定". 研究发现,将这个原理应用于语音识别是有不合适甚至是存在错误之处的.

首先,不管进行匹配的待识语音和模型是否属于同一个识别基元,算法的目标总是寻求最大似然匹配. 它没有考虑到被匹配模型与待识语音之间的关系,往往会产生错误的结果. 比如汉语"a(啊)"的语音和模型"a(啊)"、"ya(呀)"的匹配得分可能非常接近,有时甚至出现语音与本音模型的匹配得分反而不如与非本音模型的匹配得分高的情形.

其次,根据 HMM 中的马尔可夫过程的无后效性和搜索算法的特点,状态的转移仅取决于已知的累积得分和当前的状态(自该状态的转移概率和目标状态下输出特征的分布),搜索算法没有利用更多的历史知识. 有些模型也仅考虑了驻留分布[20],在某种程度上反而降低了模型对说话速度的鲁棒性.

我们认为,在进行帧同步搜索时有两种知识可以利用.

一种是基于纯统计知识的概率描述. 比较典型的是对状态驻留长度进行建模,用概率密度或分布律来刻画状态驻留长度的分布情况. 在搜索时,把系统处在当前状态的、当前驻留长度下的条件概率作为惩罚加进路径的总得分中,以此来控制 N-Best 搜索中路径的取舍[12,20].

另一种是基于统计的规则. 根据状态驻留长度的统计结果,搜索时只有驻留长度落在允许的矩形窗内的路径才可以进行状态转移或驻留. 这可以看做把第 1 种方法中的概率分布看作均匀分布. 它优于第 1 种方法,因为任何一个音的任何一个状态的驻留长度的概率分布一般都是中间密、两头疏的,但分布的两头是处于不同地位的. 当系统处于分布左侧时,我们应该控制系统尽可能地驻留在本状态;而系统处于分布的右侧时,我们应该控制系统尽可能地转移到下一个状态. 这种概率值相等情况下的状态转移倾向性的差别在第 1 种方法中被抹杀了.

本节主要讨论基于第 2 种知识的算法,称为基于统计知识的帧同步搜索(statistical knowledge based frame synchronous search,简称 SKB-FSS)算法.

(1) 利用状态驻留分布(SDD)信息

状态驻留分布(state dwell distribution,简称 SDD)是被广泛使用的、用以进行搜索剪枝的一种信息. 根据前

面的分析，我们摒弃利用驻留帧数的分布概率来进行惩罚的做法，而是对每一个状态 $s(s=0..S-1,S$ 是模型状态数目），确定一个允许的驻留帧数区间 $[D_{\min}^{(s)},D_{\max}^{(s)}]$，该区间确定了在搜索时允许的驻留长度．很显然，若仅使用 SDD，这种做法对语速变化没有很好的鲁棒性，也不能保证有很好的效果．因为区间的宽度限制了语速的变化范围，如果发音过长或过短，就不可能得到正确的识别结果．

(2) 利用相邻状态间差分状态驻留分布(DSDD)信息

考虑到状态驻留分布 SDD 对语速变化的低鲁棒性，我们不把绝对的驻留区间作为控制状态转移的参考因素，取而代之地用差分状态驻留分布(differential state dwell distribution，简称 DSDD)作为参考．对状态 $s=0$，利用较宽的驻留帧数分布 $[D_{\min}^{(0)},D_{\max}^{(0)}]$ 控制搜索时的状态驻留和转移．对状态 $s(s=1..S-1,S$ 是模型状态数目），假定允许的差分驻留帧数区间为 $[d_{\min}^{(s)},d_{\max}^{(s)}]$，已经遍历过的状态的驻留帧数代表值为 $D^{(s-1)}$，那么当前状态的有效驻留帧数区间就确定为 $[D^{(s-1)}+d_{\min}^{(s)},D^{(s-1)}+d_{\max}^{(s)}]$．

在利用 DSDD 作为控制状态转移考虑的因素之一时，由于第 0 状态允许较宽的驻留帧数分布，因此对语速变化可以有较大的鲁棒性．另一方面，由于当前状态的允许驻留区间是与历史相关的，因此，可以避免对一些没有意义的驻留帧数相对应的路径进行存储和搜索，从而提高了搜索的效率和精度．

1.3 受词法约束的词搜索树(WST)

上述的搜索算法对基元内部的搜索是有效的，但跨基元搜索时必须采取不同的方法．

在大词表(数万个词汇)的连续语音识别中，如果在搜索时采用线性的结构，效率会是非常低的，甚至难以忍受．由于汉语的词都是由音节组成的，而汉语的无调音节只有 400 多个，我们可以把词表按照音节组织成为一个有向的词法搜索网络．这个网络有一个入口结点，多个出口结点，其中任意两个结点之间的边(树枝)对应于一个音节的声学模型．有向词法搜索网络应该是效率很高的结构，但是由于每个结点的前接结点和后续结点都有可能很多，因此，其结构必然太复杂．

基于这样的考虑，我们把网络简化为一棵受词法约束的词搜索树(word search tree，简称 WST)．在这棵搜索树中，入口根结点是一个虚结点，它含有指向其子结点的存储区的信息；音节结点含有当前搜索路径的音节模型信息以及指向其子结点的存储区的信息；而叶子结点是该词的终结，是出口，它含有词的描述信息．在连续语音识别中，当路径达到叶子结点时，下一次状态转移需要重新进入这棵搜索树的根结点．

词搜索树可以提高搜索的效率，同时，相对于线性搜索，使用词搜索树可以做到以较少的路径空间去保存更多的词．利用 WST 进行词表的遍历可以看成是对线性词表搜索路径的合理剪枝，其构建和使用能够使搜索效率提高几十倍，并以有限的空间有效地保证系统性能．例如，对图 1 中的 14 个多音节词进行第 1 和第 2 音节搜索时，在 WST 中遍历，仅需保留 3 个和 7 个音节的搜索路径，而在线性搜索时各需保留 14 个音节的搜索路径．

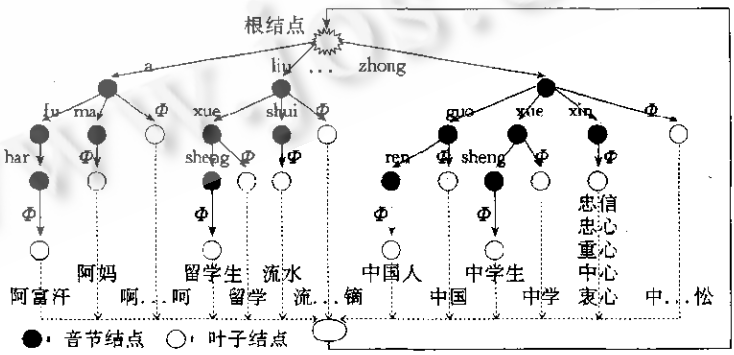

图1 一棵词搜索树(WST)示意图

1.4 连续语音识别中的双层搜索网络(TLSN)

前述的基元内和基元间的搜索构成了连续语音识别听写机系统中声学部分搜索的双层搜索网络(two-level search network，简称 TLSN)．该 TLSN 的基本结构如图 1 所示．

TLSN 的第 1 层(高层)是搜索树．该层在语音识别基元一级对搜索路径的选择进行控制，任何一个待识语

音,其搜索路径首先受搜索树的限制.比如,已知图1所示的词汇,采用音节作为识别基元,那么,当路径处在a的结束处时,下一个转移只能在fu,ma和叶子结点中选择;如果路径处在a-fu的结束处时,下一转移则只有han一个选择了.不难看出,这一层不但保证了搜索出来的词是有意义的,而且避免了不必要的搜索.

TLSN的第2层(低层)是搜索树的树枝,也就是在基元模型内部.本文提出的SKB-FSS算法在纯声学层次上保证了搜索路径的合理性和高效性,保证了搜索的效率和精度.

这种策略必须保证在搜索过程中任一时刻正确的路径不因剪枝而丢失.但是,当位于前面的音发音不好或者对应于前面音的模型训练得不好时,有可能对应正确识别结果的路径在开始时得分相对比较低,然后才会逐渐增大.但是,开始时路径得分就低,则很有可能被删掉,以后就再也无法得到正确的结果.在词表较小时,这个问题并不突出,但词表一旦增大,这个问题就不容忽视.为解决这一棘手问题,可以考虑再建立一个反向的搜索树,再进行一次自右向左的搜索,用两次结果的综合作为最后的识别结果.虽然这种方法增大了系统的时间和空间开销,但是在更大程度上保证了识别结果的正确,实践证明这也是很有效的.

在连续语音听写机中,TLSN给出的声学模型的搜索结果是一些可能的词候选列表,它们将成为语言模型的输入.因此,TLSN在语音听写机中占有非常重要的地位,其性能的好坏直接影响到语言模型的性能,从而也就影响整个听写机的性能.

1.5 语言模型

由TLSN解码得到的词候选列表,由语言模型通过语法、语义或N-Gram统计概率等加以分析,并最终给出合理的句子.语言模型有基于统计的和基于规则的两种.在基于规则的语言模型尚不成熟的情况下,N-Gram统计语言模型占有非常重要的地位.

N-Gram统计概率来自对语料的统计,但当N比较大时,无论训练语料规模多么大,也不可能包含所有有意义的N-Gram串.这样,当识别时出现训练文本中从未出现过的N-Gram串时,由于其最大似然估计为0,结果必然是错误的.也就是说,即使是对一个非常庞大的训练语料库来讲,因为数据稀疏问题的存在,最大似然概率估计也不可能对那些罕见但却是可能出现的词搭配给出一个合理的估计.

为了解决N-Gram数据稀疏问题,我们使用了一种基于Turing概率估计的方法,称为修正退化频度估计算法[21~25],并对其错误进行了改正[26].Turing概率估计方法的基本思路是:在保证概率估计总和为1的前提下,对那些出现次数大于0的N-Gram的概率估计进行折扣,并把折扣下来的值分配给出现次数为0的N-Gram,从而使出现次数为0的N-Gram的概率估计不再为0.折扣的分配并不是平均的,而是根据相应的$(N-1)$-Gram的概率大小的不同进行分配,使分配更趋合理.由于Turing概率估计法对N-Gram概率估计进行了平滑,从而降低了语言模型的困惑度(perplexity).

对一句含有M个词的语音,相对应的语音片断串为$A=A_1\ldots A_M$,声学模型对每个语音片断A_m都给出了N_m个最可能的词候选w_{mj}及其概率$P(A_m|w_{mj})$,$1\leqslant j\leqslant N_m$.如果对某个词序列$W_k=w_1^{(k)}\ldots w_n^{(k)}$,其概率为$P(A|W_k)=P(A_1|w_1^{(k)})\ldots P(A_M|w_M^{(k)})$,则语言模型的任务是在这些词候选序列中选出一个最好的词序列(也即句子):$\hat{W}=\arg\max_{W_k} P(W_k)P(A|W_k)$.

同声学模型类似,$\hat{W}$的确定也有搜索的问题.所不同的是,语言模型中的搜索是寻找最佳的Tri Gram的串接(特殊情形下使用Bi-Gram和Uni-Gram).我们尝试了最优路径和N-Best路径两种搜索策略.实验表明,最优路径搜索可以达到令人满意的结果.然而,由于语言模型不能保证完全正确,我们还利用局部N-Gram搜索进行了错误定位和纠错处理,这对降低错误率起到了很好的作用.此项研究工作尚在进行之中.

2 实验测试

为了验证上述方法,我们进行了一系列的实验.实验所用数据库是由国家863-306主题组委托中国科技大学、中国科学院声学研究所、社会科学院语言研究所等构建的汉语普通话连续语音数据库,以下简称863数据库.数据采样是在PC上利用16位标准声卡完成的,采样率为16KHz,采样精度是16bits.特征抽取的分辨率是32ms帧宽和16ms帧移.

我们确定了24 713个常用的一～四音节词作为基本词汇,其中单音节词占20.81%,双音节词占65.50%,三音节词占7.36%,四音节词占6.33%.基本词汇的选取原则是尽可能保证词汇是基本的,如"中国人民"不被收在词表中,因为它可以由没有异意的子词"中国"和"人民"串接得到.除此之外,用户可以添加定制词汇,每个定制词汇最多可以有10个音节,定制词汇的数目仅受存储的限制.

2.1 863数据库上的基本统计数据

针对863数据库,我们对音节模型中状态内驻留帧数和相邻状态间差分驻留帧数进行了统计,结果分别见表1和表2.

表1 用863数据库训练的模型中各状态驻留帧数统计(%)

驻留帧数	0	1	2	3	4	5	6	7	8
第0段	0.00	6.30	40.89	36.82	12.88	2.48	0.41	0.10	0.04
第1段	1.06	25.14	44.55	21.76	5.82	1.20	0.25	0.09	0.04
第2段	0.62	17.74	40.10	27.01	10.30	3.03	0.81	0.20	0.07
第3段	0.59	13.68	32.96	29.32	15.18	5.81	1.79	0.47	0.11
第4段	1.08	19.17	34.47	25.73	12.69	4.70	1.46	0.45	0.14
第5段	0.00	7.29	44.64	32.24	11.04	3.11	0.72	0.18	0.07

表2 用863数据库训练的模型中相邻状态间的差分驻留帧数统计(%)

差分驻留帧数	−4	−3	−2	−1	0	1	2	3	4
零一段间	0.18	2.38	14.67	35.02	32.98	12.31	2.14	0.27	0.04
一二段间	0.06	0.65	4.89	18.87	33.00	27.59	11.39	2.87	0.57
二三段间	0.20	1.27	6.25	19.06	30.41	25.53	12.28	3.80	0.92
三四段间	0.61	2.87	11.26	26.17	30.95	18.44	6.97	1.98	0.48
四五段间	0.27	1.36	5.96	20.21	35.99	27.25	7.57	1.13	0.16

可以看出,每个状态内驻留比例最大的是2帧(32ms),平均覆盖了35.0%;而比例最大的驻留帧数区间为0～5帧(0～180ms),覆盖了99.4%以上,或1～4帧(16～64ms),覆盖了95.0%以上.差分驻留帧数主要集中于0帧,平均覆盖了33%;而比例最大的差分驻留帧数区间为−4～4帧(−64～64ms),覆盖了99.9%以上,或−2～2帧(−64～64ms),覆盖了95.0%以上.

2.2 搜索算法的比较实验

在前述的24 713词的基本词表上,我们利用6状态的CDCPM音节模型以连续语音识别的方式对SKB-FSS算法进行了集外测试,其中CDN分布数目为16,使用EMM方案.从基本词表中随机选取204个词,由4个集外人录取声音数据.测试结果见表3.需要特殊说明的是,表中DSDD(Lst)是指在DSDD方案中已遍历状态驻留帧数的代表值$D^{(t-1)}$取前一状态的驻留帧数,DSDD(Avr)是指$D^{(t-1)}$取前面所有已遍历状态的平均驻留帧数.

表3 控制状态转移的搜索策略性能比较(词识别率)

策略 \ Top N		1	2	3	4	5
SDD		63.2	67.1	67.6	69.1	69.6
DSDD	(Lst)	77.9	82.8	84.8	86.8	88.2
	(Avr)	86.3	89.7	92.6	94.1	96.1

从表中可以看到,DSDD(Avr)比仅利用状态驻留信息SDD的盲目搜索的性能提高了36.6%.因此,利用不太符合语音识别的动态规划原理的帧同步算法进行"盲目搜索"不利于性能的提高,而如果把由统计知识归纳出来的一些规则(如DSDD)利用起来,进行知识导引的帧同步搜索,则可以大大提高识别器的性能和搜索效率.

DSDD规则之所以有这样好的效果,其主要原因是把一些不符合实际的搜索路径及时剪除,排除了"干扰",使得搜索算法能够把有限的搜索空间限制在有意义的路径上.

在这里,我们仅考虑了可能的搜索区间,并没有把分布的概率作为匹配得分的一部分.因为在整个有效的搜

索区间内,每一个搜索点都是合理的,不能因为统计上的差异给出不同的打分.这种考虑被实验证明是正确的.

2.3 统计语言模型的平滑效果和语言模型的组句效果

我们的语言模型以 N-Gram 为主($N=2$ 和 3),用以对语言模型进行训练的语料库包括 1993 年《人民日报》全文、市场报摘编、新华社文稿摘编共约 4 000 万词的语料,全部语料库都进行了分词和词号标注.基本词汇同上.我们用 200 个句子测试了语言模型的困惑度,对 Bi-Gram,困惑度由平滑前的 121 降为平滑后的 117,而对 Tri-Gram,则由 90 降为 88[26].

我们尝试了最优路径和 N-Best 路径两种搜索策略.实验表明,最优路径搜索可以达到令人满意的结果.我们对 863 数据库中所有 600 多个句子进行了测试,测试时对其中的词作了干扰处理(句子中每个词扩展为含有正确词的 20 个随机候选词的集合),然后利用语言模型进行句子选择,其正确率达到 94.4%.

3 汉语语音听写机引擎(CDME)处理流程

本节介绍汉语语音听写机引擎(CDM engine,CDME)内核的处理流程,该内核是使用前述技术在 Windows 95 环境下用 Microsoft Visual C++ 5.0 实现的.其处理主要分为下面的几个步骤.这些步骤并不是全部必需的,如语音命令系统就不需要"组句"过程.

(1) 循环采样

设置一个长度为 10 分钟的大循环队列,用来存放以 16KHz 采样得到的 16bit 语音数据.该缓冲区由主控进程创建的一个"采样线程"控制.采样线程连续不断地将采样加到队列的尾部;主控进程不断监测、处理(特征抽取直至识别)并释放队列的头部.这保证了用户可以连续不断地说远远超过 10 分钟的语音数据.当收到主控进程的停止采样消息时,采样线程将停止采样.主控进程继续处理队列中的剩余数据.

(2) 充分利用时域信息进行切分

对汉语普通话,人们在自然发音时,一般词与词之间存在瞬间间歇,因此,在声学搜索中加入判词信息可以把语言处理中的分词问题放在搜索中解决.本过程的任务就是充分利用各种知识(声学知识、语言知识),从循环队列的头部切分出一些相对独立而且完整的语音段,供后续过程处理.

用以切分的特征主要有能量、过零率[27]、准周期性[28]或综合倒谱信息[29]等.由各种特征,我们可以得到若干切分点,称为假想切分点(putative separation point,PSP).每个 PSP 都有一个信任度(confidence measure,简称 CM)表明对它的把握程度.对于那些超过一定信任度阈值的切点,我们定义为真实切分点(true separation point,简称 TSP),否则,称之为错误切分点(false separation point,简称 FSP).

很显然,CM 阈值的大小对 TSP 和 FSP 的分类有很大的影响.事实上,对切分点的信任度 CM 可以设置两个阈值[30,31]:接受阈值 CMA 和拒绝阈值 CMR.CM 值高于 CMA 的切分点判为 TSP;小于 CMR 的判为 FSP;而介于 CMA 和 CMR 之间的,认为是不确定的.这种信息可以有条件地利用,如只有完全确定的 TSP 才提供给后续搜索过程,此时只要把阈值选得比较苛刻,就可保证程序认定的 TSP 有接近 100%的正确率.

(3) 在确定段内进行声学层面的双层搜索,并进行接受/拒绝判决

定义相邻两个 TSP 之间的语音段为确定段 DS,它可以包含 1 个音节,也可以包含若干个音节.根据人们的说话特点,一般不会包含太多的音节数目.

这个步骤首先在确定段内进行双层搜索网络(two-layer search net,简称 TLSN)中的搜索,并给出声学词候选.然后,为减轻语言模型的负担,利用 CAP(percentage in critical area)和 RSG(recognition score gap)[31,32]参数对声学候选词进行拒识处理,向语言模型输送尽可能少而准确的候选词序列.

(4) 用语言模型进行组句

利用经 Turning 估计平滑过的 N-Gram 语言模型进行组句,并输出.

4 利用 CDME 实现的两个系统及测试结果

为了测试 CDME 的性能,我们在 Windows 95 上利用 CDME 实现了两个系统:文本编辑器 ST97 和语音命

令系统CMD97.

ST97是一个可以实际使用的"语-文转换"文本编辑器.在这个系统中,用户只要用鼠标选中了语音输入的功能,就可以一直说下去,除非再次点击停止功能.CDME将在后台不断地处理语音数据,并给出识别出来的句子.我们从863数据库中选取了100个句子进行集外测试,测试结果是词正确率达到87.6%.

CMD97是一个模拟的语音命令系统.在这个系统中,用户可以随便地定制自己的命令或地址薄,不必进行训练就可以直接口呼.我们随机选取了200个命令进行了实时(集外)测试,拒识正确率达99%,识别正确率也超过99%.

5 进一步的工作

CDME的实现向实用化迈出了一步,但我们发现其中还有很多不足和需要改进的地方.

(1) 利用特征变化量(FDS)信息

CDCPM模型的训练使用NLP算法提供状态序列解码,该算法假定"在同一状态(段)内的特征向量变化比较平稳",它保证"每段内的特征变化量总和(feature difference sum,简称FDS)大致相等",这就是等特征变化量(equal FDS,简称EFDS)的思路.NLP算法给出了在"等特征变化量"的意义下"最好的"状态序列,其优点是总能比较一致地把变化较小的那些特征子序列分到一起,从而对语速的变化有相当好的鲁棒性.

为了保证识别和训练在状态解码序列上的一致性,我们很自然地会想到在控制状态的转移时利用每个状态的FDS大致相等这样的性质.由于训练时基元的边界是已知的或准已知的,而在作连续语音识别时则是未知的,因此,控制状态转移的条件不应该是"当前状态的FDS与前一状态的FDS值相等",而应该给出可以浮动的范围.

同样,用FDS控制状态转移对语速会有较高的鲁棒性.而且,由于用FDS控制状态的驻留长度保证了识别和训练在状态序列解码上的一致性,所以性能也会有很大的改进.

(2) 声学上的鲁棒性

在声学模型方面,系统对口音的适应能力不够,这跟训练数据有很大关系.在研究中还发现,基于同态分析的倒谱特征并不是最好的特征,在特征空间没有很好的可分性.如何研究一种可分性好的声学特征是语音处理领域非常困难而重要的课题之一.

(3) 统计语言模型的后处理和规则语言模型的应用

在语言模型方面,Turing估计虽然较好地对0概率进行了平滑,而且由于其平滑是基于低阶的N-Gram的,不是平均折扣,因此取得了较好的效果.但也正因为如此,它蕴含了$[P(A_1,B_1)>P(A_2,B_2)]\wedge[P(B_1,C_1)>P(B_2,C_2)]\Rightarrow P(A_1,B_1,C_1)>P(A_2,B_2,C_2)$这种有可能是错误的推论,抹杀了在$N$-Gram中"(不可能搭配导致的)0概率与(统计不足导致的)0概率"的差别,因此有必要对0概率的N-Gram进行特殊处理.

另一方面,统计语言模型过分依赖统计概率也不太好,一些语法、语义都正确的搭配可能由于统计数据的差异导致语言模型给出错误的结果.我们的问题是:可否把一些语法、语义或语用上地位相同的一组词看成是一个等价词类,或与统计概率结合起来进行词聚类?这样既可以避免这类错误,又可以降低统计语言模型的存储规模.可否利用现有的语言处理的成果进行基于规则的语言模型研究,或者把基于统计的语言模型和基于规则的语言模型两者结合起来?这方面虽然已有一些成果,但仍是有理论和实践意义的重要研究课题.

致谢 ST97是由江门三特(SoundTek)电子科技有限公司出资、清华大学(Tsinghua)计算机科学与技术系完成的"语-文"转换(speech-to-text)系统.杨大力和张继勇同学为系统中的切分算法作出了大量的研究工作,吴文虎教授给予了大力支持和指导,在此向他们表示诚挚的谢意.

参考文献

1 Rabiner L R, Schafer R W. Digital Processing of Speech Signals. Englewood Cliffs, NJ: Prentice-Hall Inc., 1978

2 Furui S. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on ASSP, 1986,34(1):52~59

3 Juang B H, Rabiner, L R. A probabilistic distance measure for hidden Markov models. AT&T Technical Journal, 1985, 64(2):391～408

4 Ney H. Modeling and search in continuous speech recognition. In: Proceedings of the European Conference on Speech Technology. 1993. 491～498

5 Rabiner L R, Juang B H. An introduction to hidden Markov models. IEEE ASSP Magazine, 1986,3(1):4～16

6 Lee K F. Large-vocabulary speaker-independent continuous speech recognition: the SPHINX system [Ph. D. Thesis]. Carnegie-Mellon University, USA, 1988

7 郑方,武健,吴文虎等.基于最小分类错误的声学模型间距离度量.见:吴泉源,钱跃良编.第3届全国计算机智能接口与智能应用学术会议论文集(NCCIIIA).北京:电子工业出版社,1997. 98～103
(Zheng Fang, Wu Jian, Wu Wen-hu *et al*. A distance measure for acoustic models based on minimum classification error. In: Wu Quan-yuan, Qian Yue-liang eds. Proceedings of the 3rd National Conference on Computer Intelligence Interfaces and Intelligence Applications. Beijing: Electronics Industry Publishing House, 1997. 98～103)

8 Wilpon J G, Lee C H, Rabiner L R. Application of hidden Markov models for recognition of a limited set of words in unconstrained speech. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing. 1989. 254～257

9 Huang X D, Jack M A. Semi-continuous hidden Markov models for speech signals. Computer Speech and Language, 1989,(3):239～251

10 Bellegarda J R, Nahamoo D. Tied mixture continuous parameter modeling for speech recognition. IEEE Transactions on ASSP, 1990,38(12):2033～2045

11 Viterbi A J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on IT, 1967,13(2):1～18

12 Lee C H, Rabiner L R. A frame synchronous network search algorithm for connected word recognition. IEEE Transactions on ASSP, 1989,37(11):1649～1658

13 郑方,吴文虎,方棣棠.CDCPM及其在语音识别中的应用.软件学报,1996,7(10):69～75
(Zheng Fang, Wu Wen-hu, Fang Di-tang. CDCPM with its application to speech recognition. Journal of Software, 1996,7(10):69～75)

14 Zheng Fang, Chai Hai-xin, Shi Zhi-jie *et al*. A Real-world speech recognition system based on CDCPMs. In: Tang Yuan Y ed. Proceedings of the International Conference on Computer Processing of Oriental Languages (ICCPOL'97). Hong Kong: Hong Kong Baptist University Printing Press, 1997. 204～207

15 蒋力.基于概率统计模型的非特定人语音识别方法与系统的研究[硕士学位论文].清华大学,1989
(Jiang Li. The study of speaker-independent isolated word speech recognition based on statistical model [MS. Thesis]. Tsinghua University, 1989)

16 Zheng Fang, Wu Wen-hu, Fang Di-tang. Center-distance continuous probability models and the distance measure. Journal of Computer Science and Technology, 1998,13(5):426～437

17 郑方,吴文虎,方棣棠.汉语语音听写机中的语音识别基元.见:袁保宗编.第4届全国人机语音通讯学术会议(NCMMSC'96)论文集.北京:中国科学院声学研究所,1996. 32～35
(Zheng Fang, Wu Wen-hu, Fang Di-tang. The speech recognition unit in the Chinese dictation machine. In: Yuan Bao-zong ed. Proceedings of the 4th National Conference on Man-Machine Speech Communication (NCMMSC'96). Beijing: Institute of Acoustics, The Chinese Academy of Sciences, 1996. 32～35)

18 Zheng Fang, Xu Ming-xing, Wu Wen-hu. The Description of the intra-state feature space in speech recognition. In: Lee Lin-shan ed. Proceedings of the 1997 International Conference on Research on Computational Linguistics. Taipei: Academia Sinica, 1997. 272～276

19 陈永彬,王仁华.语言信号处理.合肥:中国科学技术大学出版社,1990
(Chen Yong-bin, Wang Ren-hua. Language Signal Processing. Hefei: University of Science and Technology of China Press, 1990)

20 Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of IEEE, 1989,77(2):257～285

21 Nadas A. On Turing's formula for word probabilities. IEEE Transactions on ASSP, 1985,33(6):1414～1420

22 Katz S M. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on ASSP, 1987,35(3):400～401

23 Church K W, Gale W A. A comparison of the enhanced good-Turing and deleted estimation methods for estimating probabilities of English bigrams. Computer Speech and Language, 1991,5:19～54

24 Gupta V, Lennig M, Mermelstein P. A language model for very large-vocabulary speech recognition. Computer Speech and Language, 1992,(6):331～344

25 李志敏.大字表语音识别系统中的语言模型[硕士学位论文].清华大学,1995

(Li Zhi-min. A language model in large-vocabulary speech recognition system [MS. Thesis]. Tsinghua University, 1995)

26 牟晓隆,郑方,吴文虎.基于修正退化频度估计算法的 *N*-gram 语言模型.见:王承发编.第5届全国人机语音通讯学术会议论文集(NCMMSC'98).哈尔滨:哈尔滨工业大学,1998
(Mu Xiao-long, Zheng Fang, Wu Wen-hu. A modified back-off estimation for *N*-gram based language models. In: Wang Cheng-fa ed. Proceedings of the 5th National Conference on Man-Machine Speech Communication (NCMMSC'98). Harbin: Harbin Institute of Technology, 1998)

27 郑方,吴文虎.汉语连续语音识别中音节自动切分的研究.见:张炘中编.第4届全国汉字与汉语语音识别学术会议论文集.北京:中国中文信息学会,1992. 285～289
(Zheng Fang, Wu Wen-hu. The study of automatic syllable detection of continuous Chinese speech. In: Zhang Xin-zhong ed. Proceedings of the 4th National Conference on Chinese Character and Chinese Speech Recognition. Beijing: Chinese Information Processing Society of China, 1992. 285～289)

28 黄仁忠.基音检测和汉语四声判别[硕士学位论文].清华大学,1996
(Huang Ren-zhong. Pitch detection and Chinese tone recognition [MS. Thesis]. Tsinghua University, 1996)

29 薛晓岗.汉语连续语音的切分[毕业设计论文].清华大学,1995
(Xue Xiao-gang. The segmentation of continuous Chinese speech [Undergraduate Project]. Tsinghua University, 1995)

30 郑方,胡起秀,邓翔等.介绍一种傻瓜式声控电话机.见:袁保宗编.第4届全国人机语音通讯学术会议论文集(NCMMSC'96).北京:中国科学院声学研究所,1996. 165～168
(Zheng Fang, Hu Qi-xiu, Deng Xiang *et al*. An introduction to a kind of voice dialer for dummies. In: Yuan Bao-zong ed. Proceedings of the 4th National Conference on Man Machine Speech Communication (NCMMSC'96). Beijing: Institute of Acoustics, The Chinese Academy of Sciences, 1996. 165～168)

31 郑方.连续无限制语音流中关键词识别方法的研究[博士学位论文].清华大学,1997
(Zheng Fang. Studies on approaches to keyword spotting in unconstrained continuous speech [Ph. D. Thesis]. Tsinghua University, 1997)

32 Xu Ming-xing, Zheng Fang, Wu Wen-hu. Rejection in speech recognition based on CDCPMs. In: Lee Lin-shan ed. Proceedings of the 1997 International Conference on Research on Computational Linguistics. Taipei: Academia Sinica, 1997. 412～419

Research and Implementation of the Techniques for Chinese Dictation Machines

ZHENG Fang MU Xiao-long XU Ming-xing WU Jian SONG Zhan-jiang

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

Abstract In this paper, the search strategies in the acoustic layer of the CSR (continuous speech recognition) and the CDM (Chinese dictation machine) are addressed in two aspects, the acoustic recognition unit and the syntax-constrained word search tree. The SKB-FSS (statistical knowledge based frame synchronous search) algorithm and the syntax-constrained WST (word search tree) structure are proposed, they form the TLSN (two-level search network) in the acoustic layer. The statistical knowledge used by the algorithm includes differential state dwell distribution, the feature difference sum and so on, which result in an improvement of 36.6% in CSR. The principles of a modified back-off estimation algorithm and the search algorithms for the *N*-gram based language models are also introduced. Finally, by integrating the authors' techniques, a Chinese dictation machine engine (CDME) is implemented. A speaker-independent CDM text editor named ST97 and a voice command system named CMD97 are established for personal computers (PCs) based on the CDME.

Key words CSR (Continuous speech recognition), CDM (Chinese dictation machine), search strategy, SKB-FSS (statistical knowledge based frame synchronous search) algorithm, differential state dwell, feature difference sum, WST (word search tree), TLSN (two-level search network), CDME (Chinese dictation machine engine), voice command.