邮包校核语音识别系统的实时实现

单翼翔,张昊天,李虎生,钟 林,张 进,刘 加,刘润生
(清华大学电子工程系,北京 100084)

摘 要: 本文研究开发了一套邮包信息校核语音识别系统. 该系统利用中大词汇量非特定人连续语音识别技术实时实现了邮包信息的语音校核. 系统可以识别普通话或四川话语音,可识别的词汇量约为 4500 条. 系统还采用了拒识技术与说话人自适应技术,提高了整个系统的稳健性. 实验表明对普通话的首选识别率达到 98.7 %,前三选识别率达到 99.9 %. 对四川话的首选识别率达到 95.9 %,前三选识别率达到 98.6 %,对无关语音的正确拒识率达到 85 %,对口音较重的说话人经过自适应后识别率可提高 5 - 8 个百分点.

关键词: 语音识别系统; 拒识; 说话人自适应; 信息校核
中图分类号: TN912.3 **文献标识码**: A **文章编号**: 0372-2112 (2002) 04-0544-04

The Implementation of a Practical Chinese Speech Recognition System For The Parcel Post Checking Task

SHAN Yi-xiang, ZHANG Hao-tian, LI Hu-sheng, ZHONG Lin, ZHANG Jin, LIU Jia, LIU Run-sheng
(Department of Electronic Engineerings, Tsinghua University, Beijing 100084, China)

Abstract: In this paper a real-time Parcel Post Checking Speech Recognition System is developed. This system can achieve the parcel post information checking and verification by means of a medium vocabulary speaker-independent continuous speech technology. Input speech can be Mandarin or Sichuan dialect. Vocabulary size is about 4500. By employing the advantages of the rejection model and speaker adaptation technology, the system's robustness is improved remarkably. The experiments show that the system has a very good performance. For mandarin voice input, the recognition accuracy of the top one candidate is 98.7 %, and the recognition accuracy of the top three candidates is 99.9 %; For Sichuan dialect one the recognition accuracy of the top one candidate is 95.9 %, and the recognition accuracy of the top three candidates is 98.6 %.

Key words: speech recognition; confidence measure and rejection; speaker adaptation; information checking and verification

1 引言

邮政系统在邮包的运输过程中,需要在每一个中转站对邮包是否正常到达的信息进行校核,减小运输过程出错的可能性. 通常采用的方法是在每个邮包外加上标有其发出站,目的站,以及该邮包编号的标签. 与该批邮包同时运送的还有一份相应的邮包路单. 路单上详细记录所有该批邮包的信息. 路单可以随邮包一起由押运工人携带到接收站,或以数据库文件的形式通过网络由发出站传送到接收站. 当邮包到达接收站后,邮包校核员与搬运工人一起根据路单对到站邮包逐件进行校核,确保邮包没有丢失、增加或混淆. 由于邮包的数量大,需要校核信息很多,校核工作劳动强度很大. 本文利用已经成熟的中大词汇量非特定人连续语音识别技术开发了一套邮包信息校核语音识别系统. 该系统可以替代完成原来需要靠校核员人工完成的校核任务. 校核系统安装在便携 PC 或者掌上电脑上,并在系统中载入需要校核的路单数据库文件,由校核员随身携带. 校核员通过语音输入每个邮包上的条目信息(发出站,目的站,邮包编号),系统将输入语音信息转换成文字信息并与载入系统的路单条目进行比较,同时将校核结果以合成语音回放给校核者. 校核员通过语音命令完成一系列的后继校核工作. 该系统目前在四川省成都市邮电系统中进行试用. 这不但提高了校核工作的效率和可靠性,减少了劳动强度,而且是将语音识别技术用于工业生产的一次有益的尝试.

针对该系统实际应用的需要,识别系统的输入语言可以是四川话或普通话. 根据四川话的发音特点,通过反复试验比较,本文建立了相应的发音模型. 根据路单条目信息,建立了相应的词对语言模型,基于传统的 Viterbi Beam-Search 算法采用了一种新的基于多子树结构的实时搜索算法. 由于应用环

收稿日期:2000-06-29;修回日期:2001-07-12
基金项目:国家自然科学基金(No. 69975007);国家 863 项目(No. 863-306ZD13-04-6;863-512-9805-10);中科院自动化所模式识别国家重点实验室开放课题

境背景噪声大,易引起误识,所以需要对无关语音进行拒识,本文研究开发了一种在线垃圾模型的似然度的计算方法,以及在线垃圾模型竞争集的训练算法. 考虑到校核员的口音和生理上的差异可能对系统的识别性能造成的影响. 系统还包括了语音的自适应的功能. 测试表明系统对普通话的首选识别率达到 98.9 %,前三选识别率达到 99.7 %;对四川话的首选识别率达到 95.9 %,前三选识别率达到 98.6 %;对无关语音的正确拒识率为 85 %;对口音较重的准特定人自适应后识别率可以提高 5 - 8 个百分点.

2.1　识别系统的前端处理

端点检测是实时语音识别系统中一个重要组成部分. 在本系统中采用了最大似然估值方法与首辅音能量的上升特性相结合的监测方法[7]. 之所以采用后者,是考虑到声道的运动是具有惯性的,任何语音都有一个渐变过程,不会出现类似于冲击响应的波形;而对于信道上的机械噪声或信道噪声来说,其波形往往类似于冲击响应没有渐变过程. 该方法与最大似然估值方法结合后进一步提高了端点检测的准确性,特别是在抗冲击突发噪音方面,效果很明显.

本系统采用的语音特征参数为 14 维 MEL 频标倒谱系数(Mel-frequency cepstrum coefficients),即常说的 MFCC 系数;MFCC 的 1 阶差分系数;归一化能量,以及能量的一、二阶差分组成的 31 维特征矢量.

2.2　声学模型与语言模型

声学模型采用改进的连续概率密度隐含马尔可夫模型[7]. 基本建模单位为汉语的半音节. 采用半音节模型可以减少基本识别单元并实现模型共享,这大大压缩了汉语识别中的模型存储量. 汉语的半音节可以分成两类:辅音和元音. 对于每个辅音半音节模型,采用 2 个状态的 HMM,对于每个元音半音节模型,采用 4 个状态的 HMM. HMM 模型的状态输出采用满秩的高斯密度概率分布.

在语音识别系统中,为了最大限度地压缩搜索空间,还必须根据应用的特殊性加入对应的强语法约束关系. 这种约束关系通常称为语言模型. 通过对路单条目内容的分析,实际上每句路单条目语音输入都是由 3 个关键词组合成的连续语音,即:

[邮包发送站名][邮包接收站名][邮包编号]

所以本文采用了基于约束树的多子树结构词对语法模型[8]. 首先按照语法约束信息将半音节单位以树状方式展开成搜索子网格. 然后以多子树结构为基础,展开搜索. 邮包校核系统的语法结构可以用图 2,3 来描述. 其中,邮包发送站名子树和邮包接收站名子树中的地名共享全国铁路邮政系统中常用的约 4400 条地名,而邮包编号子树中的内容除了 1—1000 的数字编号以外,还有 21 条特殊编号. 总共有 1021 条.

2　基于实用的邮包校核语音识别系统的设计[4]

邮包信息校核语音识别系统的基本结构图如图 1 所示. 本文采用的基本语音单位为汉语的半音节,对于每一个半音节采用改进隐含马尔可夫模型[7]. 采用的语言模型是基于多子树的三元词对文法模型[8]. 对于输入系统的语音命令,通过拒识模块确定该命令是否有效. 对于有效的识别结果,系统通过合成语音形式自动将识别结果回放给校核者,同时完成该语音命令需要进行的校核控制任务. 系统在实现时,采用了并行处理技术,有效地提高系统实时处理的速度.

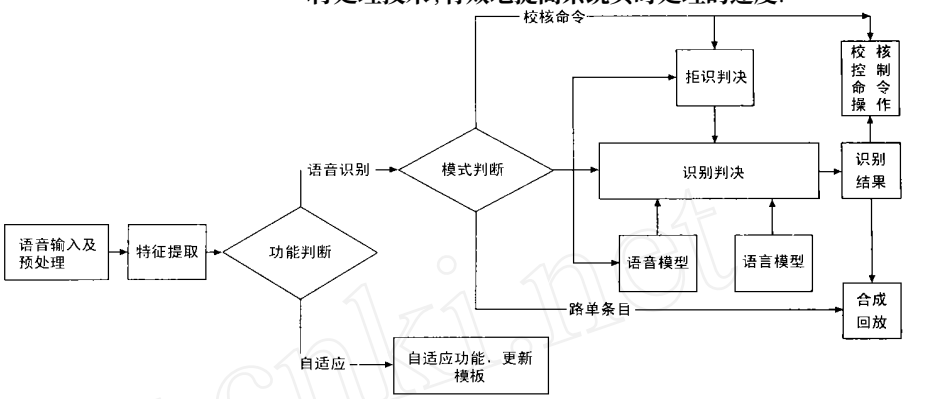

图 1　邮包信息校核语音识别系统框图

完整的路单条目集合是非常庞大的,大约有 1.977e + 10 种组合可能. 在树的生成过程中,首先统计路单记录文件中各地名及编号的出现及组合情况,以此为依据来动态生成各级子树,并建立子树之间的语法连接关系. 所以对应于每一张路单记录文件,所生成的子树和树间连接都是不同的. 但也正是这种不同恰恰代表了不同路单的数据库中相异的组合关系的信息. 从而大大压缩了原来可能的庞大的搜索空间,使得快速搜索成为可能. 同时也大大降低了信息的存储量.

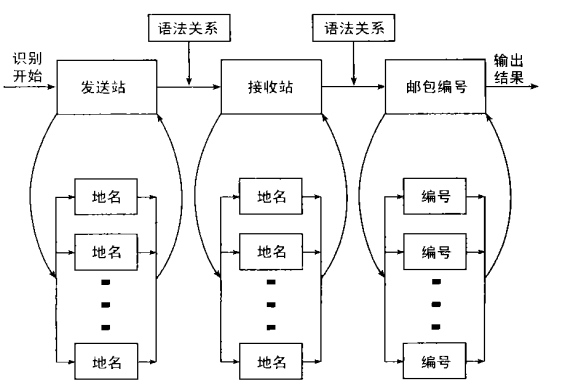

图 2　基于多子树结构的词

对一次校核任务而言,邮包条目的数量一定,而每一条校核条目都要经过确认. 那么每校核并确认一条信息,邮包条目就会减少一条,也就是说,合理的搜索路径就减少一条. 此时如果仍然沿用最初的语法信息,必然会导致搜索空间的浪费,从而降低系统的整体识别性能. 为此我们充分利用了校核前后过程中语法信息动态变化的特点,改进了子树内部的结构,

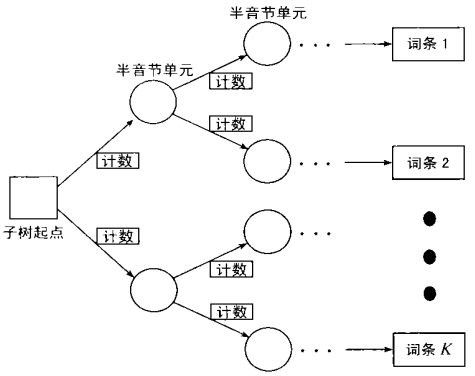

图 3 单级子树的内部结构

如图 3 所示. 利用跳转弧上的虚拟计数结构,动态地调整语法关系,并由此来指导搜索过程中的剪枝策略. 将一些已经不可能出现的路径尽早地删除,通过使用这种动态的语法修改机制,随着识别过程的继续,系统的识别率和识别速度呈现出明显上升的趋势.

2.3 训练与识别算法

对于初始化的普通话的模板,首先采用国家 863 的语音数据库作为非特定人模板的训练初始数据. 根据训练语句标注文件,将半音节模型构成复合的连续语音整词模型. 采用传统的 Baum-Welch 算法反复迭代训练. 由于在四川方言中 80 % 的语音发音与普通话相近,所以首先利用采集的四川方言语音数据对这些相应的识别模型进行细化训练;而对于剩下的 20 %左右发音有明显有差异的语音,建立专门声学模板,利用四川话方言语音数据进行训练. 将新的模型用"或"的关系[8]加入到原有的普通话模板中去,以构成混合模板. 在识别的过程中,当发现对应于某声学模型有主模板和辅助模板时,匹配的过程将在两种模板中同时进行. 匹配的分数也取两者中的高者. 在搜索的过程中,我们采用的是基于 Viterbi 算法的帧同步束搜索算法. 基本的数学模型如下:

$$\mathrm{R} = \arg\{ \min_{w} \{ l(\mathrm{O}/\mathrm{W}) \} \} \tag{1}$$

其中 $l(\mathrm{O}/\mathrm{W})$ 为声学模型提供的对数似然度, $\mathrm{W} = \{ w_1, w_2, w_3 \}$ 为识别的关键词, R 为识别出的最佳路径. 在实现的过程中,采用了非线性递减路径宽度的树搜索剪枝策略,进一步压缩了搜索空间[8].

2.4 在线垃圾模型和拒识[5]

由于在系统应用的环境中存在着各种环境噪声,同样在语音输入的过程中也不可避免地会引入各种无关的语音噪声,而这些噪声的输入往往会导致系统的误识,从而给系统带来不必要的误操作. 为避免这种情况的出现,本系统采用在线垃圾模型对无关语音进行拒识. 在线垃圾模型似然度的计算利用了分层的思想,采用了结构化的置信度计算方法,从每帧属于某个模型的置信度开始算起,依次向上计算出该语音段和识别结果中的对应词条的置信度. 当该词条的置信度小于某一门限时,系统将拒识该词条,提示用户重新输入语音. 相关置信度的计算依照以下的公式进行:

利用在线垃圾似然度 $LL^{i}_{\text{on-line}}(x)$,设 ω_i 为识别结果对应的状态模型,定义帧 x 属于第 i 个模型的置信度 $CM^{i}(x)$ 为:

$$CM^{i}(x) = \log\left\{ \frac{p(x|H_0)}{p(x|H_1)} \right\} = \log p(x|\tilde{\omega}_i) - \log\left\{ \frac{\sum_{j \neq i} p(x|\tilde{\omega}_j)}{M-1} \right\}$$
$$= LL^{i}(x) - LL^{i}_{\text{on-line}}(x) \tag{2}$$

其中 $p(x|H_0)$ 是该假设正确情况下的概率, $p(x|H_1)$ 是该假设错误情况下的概率; $LL^{i}(x)$ 为 x 隶属于第 i 个模型的似然度, $LL^{i}_{\text{on-line}}(x)$ 为对应于第 i 个模型而言, x 隶属于在线垃圾模型的似然度. 以式(2)为基础可以分别得到基于半音节和词条的置信度如下:

$$CM^{j}_{\text{semi-syllable}} = \frac{1}{N} \sum_{x} CM^{j}(x) \tag{3}$$

$$CM^{k}_{\text{words}} = \frac{1}{M} \sum_{j} CM^{j}_{\text{semi-syllable}} \tag{4}$$

$CM^{j}_{\text{semi-syllable}}$ 为语音段隶属于第 j 个半音节模型的置信度,而在最后的识别结果中共有 N 帧语音对准该半音节模型, i 为当前帧 x 对应的状态模型编号. 同样 CM^{k}_{words} 代表的是第 k 个词条,在该词条中共有 M 个半音节.

2.5 说话人自适应[6]

由于说话人生理上的差异以及训练语音数据量的不足,使得非特定人的语音识别系统在一些特定的场合下,对于某些特定的使用者来说,系统的识别率会有很明显的下降. 本系统增加了说话人自适应(Speaker Adaptation)功能来解决这个问题. 该方案利用使用者的少量训练语音,调整识别模型参数,使得系统识别性能有明显的提高. 本文采用了有监督,批处理的自适应模式. 自适应算法采用基于最大后验概率(Maximum A Posteriori, MAP)的方法. 利用 Bayes 学习理论,将原系统的先验信息与被适应人的信息相结合实现自适应. 其算法如式(5)所示. 式中 $P(\chi|\theta_i)$ 纳入了被适应人语音的信息.

$$\theta_i = \arg\max_{\theta_i} [P(\chi|\theta_i) P(\theta_i)] \tag{5}$$

其中 χ 为训练样本, θ_i 为第 i 个语音模型的参数, θ_i 为模型参数的 Bayes 估计值.

3 实验结果和讨论

在实验中语音信号的采样率为 16kHz,每个采样值用 16 比特量化. 采样信号通过 Hamming 窗处理. 帧长为 20 毫秒,帧移为 10 毫秒. 待识别地名库包含现有铁路邮政系统三级地名 4440 个,普通数字编号为 1000 个,特殊编号 21 个,共 1021 个. 普通话测试语音数据库在实验室内录制,录音的语料为成都市邮政局 1999 年 9 月 5 日从 384 次列车收到的 421 件邮包的列表. 说话人包括 5 男 3 女,其中每人读 421 句话. 四川方言训练语音数据库在成都市火车站邮包校核车间现场录制,录音的语料为全国 4100 个地名以及 0 到 999 的数字串,并按照邮包校核读法组成句子. 说话人为 39 人,均为男性,其中每人读 700 句话. 四川方言测试语音数据库在本实验室内录制,语料与普通话测试语音数据库的语料相同. 说话人为 10 名四川籍学生,均为男性,其中每人读 421 句话. 自适应测试语音库在本实验室录制,其语料与普通话测试语音数据库的语料相同,说话人为 3 人,均为男性,其中每人读 421 句话.

表 1 为邮包校核语音识别系统的测试结果. 其中 A 栏表

示基于纯普通话语音模型,由普通话测试语音数据库得出的识别性能;*B* 栏表示基于普通话与四川方言混合模型,由普通话测试语音数据库得出的识别性能;*C* 栏表示基于四川方言模型,由四川方言测试语音数据库得出的识别性能;*D* 栏表示表示基于普通话与四川方言混合模型,由四川方言测试语音数据库得出的识别性能.从中可以看出,系统的识别率达到了一个很高的水平.当然系统的整体识别率在采用了混合模板后,有一定的下降,但就识别率和实际的使用要求而言,也已经能够满足用户的需要.采用混合模板有助于减小系统开销,避免系统同时装载两套模板的存储量.采用了混合模板后,对普通话来说平均识别率仅大约只下降了 0.2 个百分点,对于四川话方言而言,识别率大约下降了 2.5 个百分点.导致该四川方言识别率下降的主要原因有两个:(1)四川方言的训练数据远远少于普通话模板的训练数据.(2)对于四川话的发音规律分析得仍然不够全面,因此四川话的模型仍然不够精细,尚待进一步完善.

表 1　使用不同模板对识别率的影响

	A	*B*	*C*	*D*
一选	98.9 %	98.7 %	98.6 %	95.9 %
二选	99.7 %	99.6 %	99.7 %	99.0 %
三选	99.7 %	99.9 %	99.9 %	98.6 %

采用说话人自适应前后,识别系统性能如表 2 所示.从表中的结果可以看出,通过自适应处理后,识别性能有了明显的改进.系统识别率最多可以提高 5～8 个百分点.

表 2　系统自适应前后识别性能比较

测试人	自适应前系统识别率	自适应后系统识别率
*T*1	83.4 %	91.2 %
*T*2	89.6 %	94.4 %
*T*3	90.2 %	93.8 %

实用的识别系统必须具有一定的拒识功能.但如果系统的拒识率过高,将会把一些本来正确的命令拒绝,导致系统使用的不友好性.相反,如果拒识率过低,将会导致无关杂音被导入系统从而引起误识,进而引起错误的操作.实验的结果表明通过在线垃圾模型在保证正确语音识别的条件下,系统能够正确拒识 85 %的无关语音.

约束语法关系的应用和合理的剪枝策略大大压缩了搜索空间.加之采用全并行处理的模式,充分利用了系统的各种软硬件资源.使得系统所具有实时处理能力.

表 3　系统实时性分析

P - 166:1.5 倍实时	PII - 350:1.03 倍实时	PIII - 600:0.7 倍实时

4　结论

本论文研究开发针对邮包校核的语音识别系统.该系统性能达到了实用的要求,已经在四川省成都市邮电局进行试用.本文就该系统实现的主要技术要点,如面向实用的中大规模词汇量连续语音识别系统的声学模型,语法模型的建立,和相应的训练识别算法进行了详细的介绍与分析.并提出了四川方言和普通话相结合的混合模型的训练方法,通过实验和实际应用结果证明了该方法的可行性.论文中还采用在线垃圾模型与说话人自适应算法提高了系统的稳健性.该系统是将语音识别技术用于工业生产的一次尝试.整个识别系统性能还有改进的余地,有关的工作还在进行当中.

参考文献:

[1] S M Ahadi, P C Woodland. Rapid speaker adaptation using model prediction [A]. IEEE Proceedings of International Conference on Acoustic Speech Signal Processing [C]. Australia: Causal Productions Pty Ltd Rundle Mall, 1995. 676 - 679.

[2] Mathan L, Miclet L. Rejection of extraneous input in speech recognition application using MLP's and the trace of HMM's [A]. IEEE Proc. ICASSP [C]. Toronto: causal Productions Pty Ltd, 1991.

[3] Colton L D. Confidence and Rejection in Automatic Speech Recognition [D]. Oregon: OHSV, 1997.

[4] 单翼翔. 邮包校核语音识别系统的实现与多线程编程 [D]. 北京:清华大学电子工程系, 2000.

[5] 钟林. 汉语语音识别说话验证 [D]. 北京:清华大学电子工程系, 2000.

[6] 李虎生. 汉语数码串语音识别及说话人自适应 [D]. 北京:清华大学电子工程系, 2000.

[7] Liu J, Jiang J T. A novel beam search algorithm of speech recognition for voice command control [J]. Chinese Journal of Electronincs, 2000, 9 (1): 56 - 60.

[8] 张昊天. 邮包校核语音识别系统的研究 [D]. 北京:清华大学电子工程系, 2000.

作者简介:

单翼翔　男,1977 年出生于上海,1999 年毕业于清华大学电子工程系,2001 年获得清华大学电子工程系网络与人机通信研究所硕士学位,主要从事语音识别,信号处理,嵌入式片上系统方面的研究.

张昊天　男,1974 年出生于山东,1997 年本科毕业于清华大学电子工程系,2000 年获清华大学电子工程系网络与人机通信研究所工学硕士学位,现在美国 CMU 攻读博士学位.研究兴趣为大词汇量连续语音识别.