语音识别片上系统中的多级搜索算法

朱 璇,陈一宁,刘 加,刘润生
(清华大学电子工程系,北京 100084)

摘 要: 本文提出了一种新的用于片上的语音识别多级搜索算法. 该算法以连续隐含马尔可夫模型(Continuous Density HMM,CDHMM)为基本识别框架. 在保证识别率基本不变的前提下,大大降低了片内存储空间的占用量,减少了识别搜索时间. 在第二级识别候选词条的选取准则上,提出一种基于置信度的选择方法,更进一步改善了识别速度,增强了识别的稳健性. 在 200 个语音命令的识别任务下,系统的识别率为 98.83 %. 而当识别词条增加到 600 条时,该算法也具有良好的识别性能.

关键词: 语音识别; 嵌入式系统; 多级搜索; 快速算法
中图分类号: TN912.3 **文献标识码**: A **文章编号**: 0372-2112 (2004) 01-0150-04

Multi-Pass Decoding Algorithm Based on a Speech Recognition Chip

ZHU Xuan, CHEN Yi-ning, LIU Jia, LIU Run-sheng
(*Dept. of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

Abstract: A novel multi-pass decoding algorithm is proposed to realize a CDHMM-based embedded speech recognition system. Compared with traditional Viterbi decoding algorithm, both the memory consumption and recognition time decreases significantly with little recognition accuracy's reduction at the same time. Furthermore, the confidence measure is adopted as the principle to determine the candidates in the second pass, which makes the recognition process more efficient. For the task of 200 voice commands, the accuracy rate achieves 98.83 %. And then, its performance can still keep steadily while the scale of recognition vocabulary is up to 600 commands.

Key words: speech recognition; embedded system; multi-pass decoding; efficient method

1 引言

近十年以来,随着大规模集成电路的不断发展,消费电子类产品,如手机、PDA(Personal Digital Assistant)等,越来越趋于小型化,以方便使用者携带. 键盘、显示屏成为了制约产品继续减小体积的瓶颈. 这样,语音就成为了一种新颖、高效的人机通信界面,语音识别则是实现这种交互方式中最为关键的技术之一. 现阶段美国[1]、德国[2]、日本[3]的一些研究机构和公司已开发出嵌入式的英语和日语语音识别系统,但是其中一些系统的识别性能仍然较差,只能应用于玩具等产品;一些系统虽然能够取得较好的识别性能,但系统的资源消耗较大,无法真正实用;而高性能的中文非特定人嵌入式语音识别算法研究在国内刚刚开始. 因此,如何在低成本、低功耗的语音识别芯片上实现高性能的非特定人汉语语音识别系统,是我们致力研究的内容. 为了在以定点 DSP 为计算核心的语音识别专用芯片上实现非特定人语音识别系统,本文提出了一种多级搜索的识别算法,充分利用了 DSP 的数据和程序的重载功能,降低了系统对于片内 RAM 的需求,提高了识别速度,同时仍然保持了较高的孤立词识别率.

2 多级搜索算法基本原理

2.1 基线系统

基线系统的语音采样频率为 16kHz,A/D 采用 16 位线性量化器. 输入特征矢量为 31 维,其中包括 14 维的 Mel 频标倒谱系数,14 维的一阶差分 MFCC 系数,归一化能量及其一阶和二阶差分[4]. 该系统以基于音素的 CDHMM 为基本识别框架,构成嵌入式中等汇量语音命令识别系统. 采用了字内上下文相关的音子模型,聚类以后,总计 788 个状态. 观察矢量对应于每一个状态的概率分布为混合高斯(Gaussian Mixture)密度函数,其高斯混合分量数为 6,每一个高斯混合分量的协方差矩阵均为对角阵[5]. 而基线系统的搜索算法采用了基于 Viterbi 解码的帧同步搜索.

根据 Viterbi 解码算法可知,语音识别阶段的主要计算量有两个方面:观察概率的计算和网络的搜索. 由于在该系统为孤立词识别,任务相对简单,识别网络还作了相应的简化,因此网络搜索资源消耗量非常低. 另一方面,为了达到良好的识别性能,必须采用比较复杂的混合高斯分量作为观察概率估值,因此观察概率的计算占用了绝大部分的识别时间. 此外,

收稿日期:2002-08-15;修回日期:2003-01-04
基金项目:国家自然科学基金(No. 60272016)

由于基本音子数量有限,同一音子在识别网络当中往往多次重复出现.为了避免重复计算,在实用系统中将专门开辟一段内存,用来存储每一帧语音对应每一个 HMM 模型的观察概率,其为基线系统中最主要的内存消耗部分.事实证明,当基线系统采用了上文所述的音子划分和 HMM 模型,识别词条数为 200,词条所包含的平均字数为 4,最大字数为 6,待识别语音的长度不超过 200 帧(即 3.2 秒),DSP 的运算速度为 80MIPS 的情况下,硬件资源的消耗主要包括:网络搜索的时间大约为 0.1 秒,内存占用量为 12k 字节;而计算观察概率的时间约为 3.3 秒,内存占用量为 350k 字节.显然,为了实用考虑,该系统还需要更进一步降低资源消耗.

2.2　基本原理

由上文可知,基线系统无法在片上得到实现,其主要的瓶颈在于过长的识别时间,特别是计算系统的观察概率所消耗的时间,通常解决这一问题的经典方案是采用束搜索算法.但是,束搜索虽然能够在一定程度上解决计算时间过长的问题,他仍然存在以下几个问题:(1)为了达到束搜索的目的,增加了排序和比较指令,导致网络搜索时间的增长;(2)为了在良好的识别率和较快的识别时间当中取得一个平衡点,稳健的束搜索宽度设置一直是搜索算法中一个难点.

而在中等词汇量的情况下,片上语音识别系统一般包含了几百个识别词条.在较好构造声学模型情况下,绝大多数的错误识别结果的识别似然度,往往同正确结果的似然度相差甚远.为了将这些同正确识别结果相距甚远的词条尽早剔除出识别过程,本文提出一种基于声学层模型的新型多级快速搜索算法.

在第一级识别过程当中,采用了比基线系统更为简单的声学模型,其状态数量简化为 356 个,而每一个观察概率分布函数的高斯混合分量数为 1,其协方差矩阵为对角阵.虽然在识别任务比较复杂时,采用这种简单的识别模型会导致系统无法得到准确的识别结果,但是,待识别语音同正确识别模板之间的匹配分数仍然较高.采用简化模型的前 5 选的识别结果正确率可达到 99%以上,而前 10 选的正确率可以达到 99.68%.如果在此基础上,采用基线系统中的模型进行第二级识别,可以得到同基线系统相当的识别率,两级搜索算法的程序基本流程如图 1 所示.

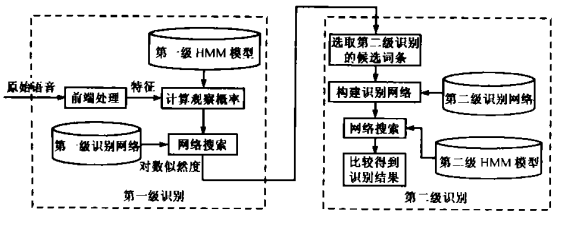

图 1　两级搜索算法的程度基本流程

在两级搜索的过程中,由于在第一级识别采用了比较简单的音子模型和 HMM 模型.这样,存放观察概率的内存空间从基线系统中的 350k 字节减少到 230k 字节,而网络搜索所占用的内存空间仍然为 12k 左右,第一级识别的最大消耗时间为 0.25 秒.

在第二级识别当中,根据第一级识别中待识别语音和各词条网络匹配得到的对数似然度(Logarithmic Likelihood,LL),选择第二级识别的候选词条.再通过比较复杂的音子模型和 HMM 模型进行解码,可以得到更为准确的识别结果.由于第二级识别的内存占用量比较小,因此,两级搜索算法的内存占用量,完全取决于第一级识别的内存占用量.在时间消耗上,当第二级候选词条的平均长度为 4 字时,每一个词条的识别时间约为 0.1 秒.由于,第二级识别的识别时间同第二级的候选词条数量成正比,因此,为了避免过长的识别时间,需要对第二级识别的候选词条数量设置上限.

在两级识别的过程当中,由于基本的搜索方法仍然是 Viterbi 解码算法,因此,各种经典的快速算法,均能直接运用于多级搜索的两个阶段中,进一步加速识别的过程.而第一级识别和第二级识别的过程相对独立,这使得语音识别芯片的重载功能得以充分的利用,在更低的资源消耗下,达到比基线系统更优的整体性能.

2.3　基于置信度的第二级候选词条选取准则

正如上文所述,两级搜索算法的第一级识别过程的内存占用量和解码时间相对比较固定,而第二级识别的识别时间同第二级候选词条数量成正比,因此,选择合适的候选词条是提高两级搜索系统整体性能的关键,而最直接的方案就是选取似然度最高的前 V 个词条作为第二级识别的候选词条.但是,这种选取准则实际上是假设正确识别结果的候选数的分布为均匀分布,而事实并非如此.因此二级识别应该根据第一级识别所得到的对数似然度信息,计算第一级识别结果的置信度(Confidence Measure,CM),来选择第二级识别的候选词条.

本文中置信度的计算采用了在线垃圾模型[6](Online Garbage Model,OGM)和归一化对数似然比(Normalized Log Likelihood,NLL)的线性组合,即:

$$CM = \lambda \times CM_{OGM} + (1 - \lambda) \times CM_{NLL} \tag{1}$$

其中词条 v 的在线垃圾模型置信度为:

$$CM_{OGM}(v) = \log \frac{\exp(LL_v)}{\sum_{k=1}^{K}\exp(LL_k)} = LL_v - \log \sum_{k=1}^{K}\exp(LL_k) \tag{2}$$

LL_k 为词条 k 的对数似然度.而词条 v 的归一化对数似然比置信度为:

$$CM_{NLL}(v) = \frac{1}{\text{FrameNum}}(LL_{\max} - LL_v) \tag{3}$$

其中 FrameNum 为待识别语音的帧数.

完成第一级识别以后,计算各词条的置信测度 $CM(v)$.其中,λ 为两种置信度之间的比例系数,实验表明,当 $\lambda = 0.1$ 时,系统能够达到最优的性能.在 $\lambda = 0.1$ 一定的情况下,只有满足 $CM(v) \geq Th$ 的词条,为第二级识别的候选词条.选择不同的阈值 Th,可以根据任务的需要,调整系统的整体性能.

3　实验结果比较

系统的训练集为 863 大词汇量连续语音数据库,本文采用其中男声数据,包括 83 人的录音,总计 46 小时的语音,语音的采样率为 16kHz,16 位线性量化,录音环境为安静的办公

室,语音语料均选自 93、94 年《人民日报》,朗读方式发音. 系统的识别集为本实验室自录的孤立词库,10 名男性,每人念 600 个词条,包括 200 个地名、200 个人名、200 个股票名称. 录音环境为办公室环境,16kHz 采样,16 位线性量化.

实验一 指定第二级候选词条数量的选取准则下,系统的识别性能. 系统测试词表的词条数量为 200,第二级识别的候选数量 $V = 1 \sim 10$. 系统的识别率和识别时间如表 1 所示,其中,系统的识别时间以基线系统的识别时间为基准.

采用多级搜索算法,能够用较小的数据 RAM 占用量(大约 250k 字节)和较少的计算时间(10 候选时的峰值识别时间为 1.3 秒)得到和基线系统相似的识别率. 但是,即使在候选词条数量已经为 10 的情况下,系统的第一级识别率仍然无法达到 100 %,而两级搜索的识别率为 98.45 %,比基线系统的 98.91 %大约低了 0.5 个百分点.

表 1 指定第二级候选词条数量的系统性能

基线系统	识别率			识别时间
	98.91 %			100 %
多级搜索系统	候选数量	第一级多候选识别率	第二级识别率	识别时间
	1	95.70 %	95.70 %	9.23 %
	2	98.17 %	97.23 %	12.98 %
	3	98.82 %	97.94 %	16.15 %
	4	99.15 %	98.16 %	18.87 %
	5	99.38 %	98.38 %	22.03 %
	6	99.48 %	98.40 %	24.84 %
	7	99.55 %	98.42 %	27.67 %
	8	99.57 %	98.42 %	30.79 %
	9	99.60 %	98.45 %	33.69 %
	10	99.67 %	98.45 %	36.87 %

实验二 不同的阈值 Th 下,多级搜索系统的识别性能. 以第一级识别结果的置信测度 $CM(v)$ 为选择第二级识别候选词条的准则时,阈值 Th 会影响系统的识别率和识别时间. 表 2 显示了 Th 变化时,系统的识别性能. 其中,识别词表的词条数为 200,系统的识别时间仍然以基线系统为基准.

表 2 采用不同阈值的多级搜索系统识别性能

阈 值	第一级多候选识别率	第二级识别率	平均候选数	识别时间
- 0.25	97.56 %	96.92 %	1.13	10.52 %
- 0.50	98.55 %	97.99 %	1.26	11.59 %
- 0.75	99.38 %	98.50 %	1.82	13.34 %
- 1.00	99.68 %	98.83 %	2.61	18.41 %
- 1.25	99.76 %	98.86 %	3.83	23.49 %
- 1.50	99.85 %	98.87 %	5.69	29.78 %
- 1.75	99.89 %	98.88 %	8.35	41.83 %
- 2.00	99.97 %	98.88 %	11.79	52.27 %

可见,当 $Th = -1.00$ 时,系统的最后的识别率可以达到 98.83 %,同基线系统的识别性能相当. 这时第一级识别提供的平均候选数仅为 2.61,第二级识别平均仅需要从第一级识别结果中不到 3 个候选中确认一个识别结果,而总的识别时间仅为基线系统的 18.41 %. 较之指定候选数量的两级识别方法,其识别精度优于采用 10 候选的情况,而识别时间则低于 4 候选. 显然,以第一级识别结果的置信度作为选择第二级识别候选词条的准则,能够以更少的消耗,得到更优的效果.

实验三 取 $Th = -1.00$,改变识别词表词条数量比较系统的识别性能. 表 3 比较了词表词条数量分别为 600、200、100、50、20、10 时,基线系统和多级搜索系统的识别率.

表 3 系统对于词表词条数量的适应性比较

词条数量	基线系统识别率	多级搜索识别率	平均候选数量
600	97.90 %	97.73 %	3.83
200	98.91 %	98.83 %	2.61
100	99.12 %	98.97 %	1.89
50	99.47 %	99.45 %	1.43
20	99.73 %	99.73 %	1.17
10	99.87 %	99.87 %	1.09

从表 3 中的识别结果可以看出:在不调整 Th 的值情况下,对不同的词表词条系统仍然可以获得同基线系统相当的识别率. 而且,随着词条数量的减小,第二级识别的词条候选数量也会相应的减少,当词条数量为 10 时,第二级识别的平均候选数量为 1.09,第二级识别率与第一级识别率相同.

4 结论

本文研究了嵌入式非特定人语音识别系统的识别算法. 首先,描述了基于传统 Viterbi 算法的基线系统,并且指出运算时间过长是制约该系统在片上实时实现的主要障碍. 针对该问题,提出了一种新的多级搜索的识别算法,在第二级的候选词条的选择上,采用了一种新颖的置信度准则. 当系统词表为 200 条词条时,系统的峰值内存占用量为基线系统的 66 %,识别时间为基线系统的 18.41 %,而识别率仅从 98.91 %降低为 98.83 %. 而当系统的词表规模达到 600 条时,系统的识别率仍然能够保持很高的识别性能.

致谢 德国 Infineon 公司为本算法研究提供了核心处理芯片及其开发工具,在此表示诚挚的谢意!

参考文献:

[1] Gong Y F, Kao Y H. Implementing a high accuracy speaker-independent continuous speech recognizer on a fixed-point DSP[A]. Proceedings of ICASSP[C]. Piscataway: Institute of Electrical and Electronics Engineers Inc, 2002. 3686 - 3689.

[2] Burchard B, Romer R, Fox O. A single chip phoneme based HMM

speech recognition system for consumer applications[J]. IEEE Trans on Consumer Electronics,2000,46(3):914 - 919.

[3] Menendez-Pidal X,Duan L,Lu J W,et al. Efficient phone based recognition engines for Chinese and english isolated command applications [A]. Proceedings of International Symposium on Chinese Spoken Language Processing[C]. Taipei:Institule of China Computational Linguistics,2002. 83 - 86.

[4] Zhu X,Wang R,Chen Y N,et al. Acoustic model comparison for an embedded phoneme-based mandarin name dialing system[A]. Proceedings of International Symposium on Chinese Spoken Language Processing[C]. Taipei:Institute of China Compatational Linguistics,2002. 9 - 12.

[5] 杨行峻,迟惠生. 语音信号数字处理[M]. 北京:电子工业出版社,1995.

[6] Bourlard H,D hoore B,Boite J M. Optimization recognition and rejection performance in wordspotting systems [A]. Proceedings of ICASSP [C]. Piscatariay:Institate of Electrical and Electronics Engineers Inc. 1994. 373 - 376.

作者简介:

朱　璇　女,1977 年 4 月生于湖北省武汉市,1998 年 7 月毕业于北京航空航天大学电子工程系,获学士学位,1998 年 9 月进入清华大学直接攻读博士学位,现为清华大学电子工程系博士研究生,主要研究方向:嵌入式语音识别系统的研究.

陈一宁　男,1976 年 8 月生于北京市,1999 年 7 月毕业于清华大学电子工程系,获学士学位,1999 年 9 月继续直接攻读博士学位,现为清华大学电子工程系博士研究生,主要研究方向:口语对话系统的研究.

刘　加　男,1954 年 12 月生于福建省福州市,1983 年 7 月获得清华大学电子工程系无线电技术专业学士学位,1986 年 7 月获得清华大学电子工程系通信与电子系统硕士学位,1990 年 4 月获得清华大学电子工程系通信与电子系统博士学位,1990～1992 年中科院遥感卫星地面站从事美国 6 号陆地卫星图像处理系统开发工作,1992～1994 年在英国剑桥大学作博士后,从事语音识别系统研究工作,现任清华大学电子系教授,博士生导师,中国电子学会高级会员,美国 IEEE 会员,近 5 年来在国内外期刊及学术会议上发表论文 40 多篇,目前研究方向包括:语音识别、语音合成、语音编码、语音识别专用芯片设计,多传感器融合技术,以及多媒体数字通信系统.

刘润生　男,1933 年 9 月生于河北省黄骅市,1958 年清华大学无线电专业毕业留校任教至今,现任清华大学电子工程系教授,博士生导师,长期从事脉冲数字电路、模拟电路、集成电路、集成电路设计、电子电路 CAD、通信与信号处理等方面的教学和科研工作,近十年来在国内外期刊及学术会议上发表论文八十余篇,其中在核心刊物以及重要国际会议上发表并为 EI 收录的论文三十余篇,目前的主要研究方向包括:汉语数码及中小词汇量语音识别及其应用,数字信号处理及模数混合集成电路设计,电子电路快速模拟算法的研究.