文章编号:1008-0570(2008)06-2-0004-02

嵌入式英语命令词语音识别算法研究

Embedded English Speech Commands Recognition Algorithm Research

(清华大学) 姚　竞　王国梁　刘　加
YAO Jing　WANG Guo-liang　LIU Jia

摘要: 本文提出了一种基于定点 DSP 的嵌入式英语语音命令词识别算法,并基于 TI 芯片建立识别系统。系统采用基于连续隐 Markov 模型(Continuous Density Hidden Markov Model, CDHMM)的两阶段识别策略。通过决策树结合数据驱动的状态聚类方法,一阶段模型数目研究等方法提高识别率。最后在以 TI TMS320vc5502 定点 DSP 为核心的语音处理片上系统上实现了英语语音命令识别,当 DSP 工作速度为 200MIPs 时,实时率为 0.37,存储空间消耗为 49.5kbyte,对于 1235 词的识别效果为 95.4%。

关键词: 语音识别;嵌入式;状态共享;特征选择
中图分类号: TN912.3　　**文献标识码**: A

Abstract: This paper presents an embedded English commands speech recognition system based on fixed-point DSP and establishes recognition system based on TI chip. The system uses Continuous Density Hidden Markov Model (CDHMM) and two-pass search strategy. To increase the recognition accuracy, phone models tying based on decision tree and data driven, phone model unit choosing in first stage recognition are applied in the system. The English commands speech recognition system has been realized on speech processing system on chip based on TI TMS320vc5502 fixed point DSP. When the DSP works on speed of 200MIPs, the system can provide a recognition accuracy rate of 95.4% on 1235 two-word phrases recognition task with 0.46 times lower than real time and 49.5kbyte data space.

Key words: Speech recognition; Embedded system; State tying; Feature selection

技术创新

1 引言

近年来,随着电子技术的发展,智能式移动设备在实际生活中得到了越来越广泛的使用,在实际应用中,迫切需要更快捷、方便和小型化的人机交互界面,而传统的小键盘或触摸盘设备在这方面并不能给出理想的答案。与此同时,自动语音识别技术(ASR, Automatic Speech Recognition)技术也得到了迅速的发展,一些简单的语音识别系统,已经可以应用到嵌入式平台上,例如电话语音拨号,智能玩具,机器人控制等。由于语言本身是人类最常使用的交流方式,因而,嵌入式语音识别技术,必将成为今后智能式移动设备用于人机交互的一项重要选择。

目前的嵌入式非特定人语音识别系统,以基于 CDHMM (Continuous Density Hidden Markov Model)的识别方法为主。但由于嵌入式平台本身运算能力和存储容量的限制,往往不能在识别性能和硬件资源上取得比较好的平衡。或者由于识别性能差而使应用场合受到限制,或者由于追求识别性能而使硬件资源激增。如何在有限的嵌入式平台上实现反应速度快,资源消耗量少,识别性能高的语音识别系统,是我们所关心的内容。

本文提出了一种基于二级搜索的识别策略,并针对嵌入式语音识别特点对声学模型进行改进,从而在以定点 DSP 为核心的嵌入式平台上,实现了中等词汇量的英语命令词语音识别系统,具有识别性能高,资源消耗量少的特点。

姚　竞: 硕士研究生
基金项目:国家自然科学基金资助项目(60572083)

2 基线系统

2.1 基线系统

语音识别基线系统如图 1 所示,其中特征提取采用 39 维的 Mel 频标倒谱系数 (Mel Frequency Cepstrum Coefficient, MFCC),识别采用基于音素的 CDHMM 模型作为基本框架,搜索算法采用了基于 Viterbi 解码的帧同步搜索。

2.2 识别策略

为了解决高识别率、复杂模型与有限的硬件资源之间的矛盾,本文采用了一种二级搜索算法。如图 1 所示,在一阶段识别中,采用简化 Triphone 模型和静态识别网络进行快速识别,得到多候选识别词条,在二阶段识别中,利用一阶段得到的多候选词条构建网络,采用精确 Triphone 模型进行识别,得到最后的识别结果。采用这种方法,由于在一阶段中模型复杂度较低,从而节省了硬件资源,提高了识别速度。而在二阶段时,待识别的词条数已经很少,从而可以使用精确 Triphone 模型得到高识别性能。

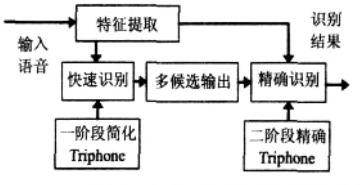

图 1 语音识别两级搜索算法

3 声学模型的改进

3.1 训练集与测试集

本文所采用的语音数据训练集为 LDC WSJ1 训练库(SI_TR_S),包括 200 人的连续语音,共 61 hours,用低通滤波器将采样率从原始的 16 kHz 降为 8 kHz,16bit 量化。

测试集为由 WSJ1 测试集(CDTest 和 HSDTest)得到的 3 组 1255 个短句(每句包含 2 个单词),每组候选词条为 1235 个。

3.2 Triphone 模型的状态聚类

在英语发音过程中,由于音素之间的变音情况比较多,因而采用 Triphone 模型,能够更准确地反映英语音素之间的协同发音,取得更好的识别效果。但由于 Triphone 模型数量众多,对于 n 个音素,需要 n^3 个 Triphone 模型来描述,如此复杂的模型在实际应用中是无法承受的,因而采用一定的方法进行状态聚类是必需的。

在英语发音过程中,由于音素之间的变音情况比较多,因而采用 Triphone 模型,能够更准确地反映英语音素之间的协同发音,取得更好的识别效果。但由于 Triphone 模型数量众多,对于 n 个音素,需要 n^3 个 Triphone 模型来描述,如此复杂的模型在实际应用中是无法承受的,因而采用一定的方法进行状态聚类是必需的。

在本文中,采用决策树(Decision Tree,DT)聚类与数据驱动(Data-Driven,DD)相结合的方法进行状态聚类,设计 Triphone 模型结构为 3 状态,输出概率密度函数为 8 mixture 的 GMM(Gaussian Mixture Model)模型。利用决策树聚类方法聚类得到 1635 个模型,然后利用数据驱动方法聚为 674 个,组成第二阶段的识别模型。

在数据驱动聚类过程中,我们定义两个 Triphone 模型 i 和模型 j 的散度距离如下:

$$d(i,j)=\log\left[\sum_{m=1}^{M} b_{jm} N\left(\mu_{im};\mu_{jm};\Sigma_{jm}\right)\right]+\log\left[\sum_{m=1}^{M} b_{im} N\left(\mu_{jm};\mu_{im};\Sigma_{im}\right)\right] \quad (1)$$

$$N\left(\mu_{im};\mu_{jm};\Sigma_{jm}\right)=\frac{1}{\sqrt{(2\pi)^n\left|\Sigma_{jm}\right|}}\exp\left[-\frac{1}{2}\left(\mu_{im}-\mu_{jm}\right)'\Sigma_{jm}^{-1}\left(\mu_{im}-\mu_{jm}\right)\right] \quad (2)$$

其中:M 表示 GMM 状态模型的 Mixture 个数;b 表示各 Mixture 分量的权重;μ 和 $|\Sigma_{jm}|$ 分别表示 GMM 模型中的均值和协方差矩阵。

实验结果如表 1 所示,测试了采用 DT+DD 方法得到的 Triphone 模型的识别率,并与单纯 DT 方法得到的 Triphone 模型进行对比,两种模型均为 680 状态左右,结果说明采用 DT+DD 方法进行聚类识别性能得到明显改善。

表 1 DT+DD 聚类方法模型识别率

DT		DT+DD	
状态数	识别率(1 选)	状态数	识别率(1 选)
690	95.6%	674	96.4%

3.3 第一阶段音素单元选择

由于一阶段属于静态识别,模型数目与复杂程度均受到 DSP 硬件资源的限制,只能采用比较简单的音素单元体系,例如 Monophone。在基线系统中,采用了 CMU 音素体系,包含 40 个音素。但实验证明,Monophone 模型状态数量过少,在进行一阶段识别时性能较差,需要通过增加更多候选词条数,直接影响第二阶段识别的识别性能和资源消耗。

本文中,对于一阶段识别采用一个简化的 Triphone 模型系统进行识别,利用精确 Triphone 模型采用数据驱动的聚类方法得到 209 状态的简化模型,实验结果如表 2 所示,从结果看,识别效果比 Monophone 模型有了提高,并可大大减少多候选识别词条数。从而充分利用了系统的硬件资源,提高了识别效果。

4 嵌入式语音识别系统的实现

4.1 硬件平台

本语音识别系统的硬件平台是以 TI TMS320VC5502 定点 DSP 芯片为核心,结合 AD,DA,FLASH,串口通信等组成的语音处理片上系统,如图 2 所示。TI 5502 DSP 是 16bit 定点 DSP,最高可以达到 300Mips 的运算速度,拥有 64k byte 的片内缓存。在片上系统中,在 DSP 周围外接了 2M NORFLASH,双路 16bit AD DA 通道等其他外部接口和设备。

表 2 一阶段识别效果对比

候选词条数目	模型识别率	
	Monophone	Triphone
1	88.7%	90.6%
2	94.0%	95.6%
3	95.4%	96.8%
4	96.2%	97.8%
5	96.7%	98.2%
6	97.0%	98.5%

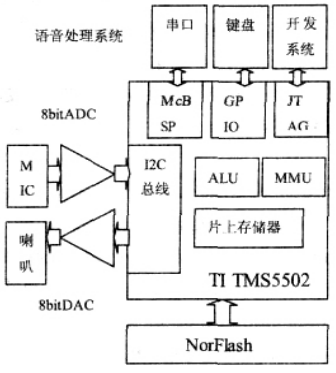

图 2 TI5502 板级语音处理系统

4.2 系统实现

本文在 Ti 5502 定点 DSP 上实现英语命令词语音识别的过程中,对 PC 机上的浮点识别程序进行了定点化,并进行了适当的代码优化,以提高代码的执行效率,满足嵌入式平台的硬件资源要求。表 3 统计了利用 Ti TMSVC5502 DSP 芯片进行 1000 词英语识别各部分所占用的系统资源。当 DSP 工作在 200MIPs 的运算速度上时系统实时率为 0.37,内存资源占用量为 49.5kb。定点化后,片上语音识别系统对于 WSJ1 测试集的两阶段识别率为 95.4%。

表 3 系统资源消耗表

主要识别流程	运算速度(实时率)	数据空间
特征提取	0.047	2.8kb
第一阶段识别	0.245	40.7kb
第二阶段识别	0.04	21.0kb
语音识别系统	0.37	49.5kb

5 结论

本文给出了一种基于定点 DSP 的嵌入式英语语音命令识别系统,采用两阶段 CDHMM 模型进行识别,第一阶段采用 209 状态 triphone 模型进行快速识别,第二阶段采用 674 状态 triphone 模型进行精确识别。通过模型状态共享和模型数目研究等方法提高识别率。并通过定点化和识别策略改进等方法在嵌入式平台上降低硬件资源消耗,最后在以 TI TMS320VC5502 定点 DSP 为核心的语音处理片上系统上实现了英文命令词语音识别系统并进行了测试,当 DSP 工作速度为 200MIPs,实时率为 0.37,存储空间消耗为 49.5kbyte,对于 1235 词的片上识别效果为 95.4%。

(下转第 8 页)

技术创新

空间,访问相应的存储空间时,就会产生一个片选信号。本系统中,将 AD2S83 视作一个外设,映射到大小为 8k×16 的存储空间 XINTF Zone0,与 Zone1 共用一个片选信号:/XZCS0AND1。由于只有 AD2S83 一个外设,所以不需要区分 Zone0 和 Zone1。在使用外部存储器接口时,需要根据 F2812 器件的工作频率以及 XINTF 特性进行参数配置,配置程序如下:

```
XINTCNF2[0] = XINTCNF2[0]&0xfff8;    //XCLKOUT=XTIMCLK,没有写缓冲
XINTCNF2[1] = 0x0001;        //XTIMCLK=SYSCLKOUT/2
XTIMING0[0] = 0x3FFF;    //建立、激活、跟踪周期均为最大
XTIMING0[1] = 0x0043;        //X2TIMING=1
```

任何对 XINTF 空间的读或写操作的时序都可以分为三个阶段:建立、激活和跟踪。建立阶段所访问空间的片选信号为低电平,产生有效的地址在 AB 上;激活阶段读选通信号/XRD 变为低电平,数据锁存到 DSP;跟踪阶段则在读信号变为高之后,保持片选信号低电平一段时间。

AD2S83 上通过/INHIBIT 和/ENABLE 两个输入信号来控制数据总线上数据的有效性,在/INHIBIT 被置低 490ns 之后数据才会有效,所以在程序设计中需要根据所使用 DSP 晶振的不同选择合适的 SYSCLKOUT 以保证时序上的一致性。

由于扩展存储空间只用到 Zone0, 所以将空间 Zone0 直接定义为寄存器,直接通过该寄存器来读取数据:

```
volatile unsigned int *XINTFZone0=(volatile unsigned int *) 0x002000;  //外部扩展存储空间 0 地址 0x0000-2000
```

当然,可以通过 FPGA 进行地址译码,这样可以在多个外设的情况下不会互相影响,而且可以实现多个外设映射到同一个存储空间的不同地址段。

5 结语

本系统使用 DSP 以及专用于旋转变压器和感应同步器的 R/D 转换器 AD2S83 实现了一种跟踪型感应同步器数字测角系统,直接转换为数字量,便于实现复杂控制算法,并且经过实验具有良好的测量稳定性。由于 DSP 作为现在主要的数字信号处理器,可以方便地实现对系统的各种控制算法,并且自带串口、CAN 等模块,加上 FPGA 作为逻辑控制器件,能够实现更加复杂的计算机接口,整个设计具有良好的可扩展性。

本文作者创新点:实现了基于 DSP 和 AD2S83 的数字式感应同步器测角系统。通过激磁电路采取单相激磁鉴相的方式,将双相输出信号输入基于跟踪型感应同步器测角系统 R/D 转换芯片 AD2S83,并通过 FPGA 选择其分辨率,在不同分辨率下使用四路通道开关选择外围器件,最终输出数字位置和转向信号,具有较好的稳定性。

参考文献

[1] 彭俊峰等.三种轴角数字转换电路的分析与比较[J]微计算机信息,2006.No.22: 8-10

[2] AD 公司.《Variable Resolution, Resolver-to-Digital Converter AD2S83》.Analog Devices, Inc., 2000

[3] 孙力等.跟踪型感应同步器测角系统特性分析.微特电机,No.5: 20-22,1996

[4] 端木时夏等. 感应同步器及其数显技术》.上海:同济大学出版社,1990

作者简介:严春晓(1984-),男,汉族,江苏南通人,在读硕士研究生,研究方向为导航与自动控制; 张嵘(1969-),男,汉族,湖北人,清华大学精密仪器与机械学系副研究员。

Biography:YAN Chun-xiao(1984-), male, postgraduate student in Department of Precision Instruments and Mechanology of Tsinghua University, research direction: navigation and automation.

(100084 清华大学精密仪器与机械学系) 严春晓 张 嵘

通讯地址:(100084 北京市 北京市海淀区清华大学精密仪器与机械学系 9003 大楼导航技术工程中心) 严春晓

(收稿日期:2008.4.05(修稿日期:2008.5.22)

(上接第 5 页)

本文作者创新点:1. 本文给出了一种基于目前主流 DSP 核心的嵌入式英语命令词语音识别系统,并在实际硬件系统上得到了实现,具有识别率高,硬件消耗资源少等特点,可以广泛应用于各种智能移动平台的语音识别任务;2. 本文在英语命令词语音识别系统的片上实现中,提出逐段匹配结合逐段网络搜索的识别策略,该识别策略可在硬件资源受限的情况下进一步降低资源占用量,从而扩大了该语音识别系统的适用范围。

参考文献

[1] 吴智量, 陈智昌, 陈烘华, 黄镜洪 语音识别控制在音频视频系统中的应用[J]微计算机信息 2004.1-3 :p113-114

[2] 黄涛, 胡宾 基于 SPCE061A 单片机的非特定人语音识别设计[J]微计算机信息 2006,3-2:p19-20

[3] Novak M, Hampl R, Krbec P, et al. Two-pass search strategy for large list recognition on embedded speech recognition platforms [C]. Proceedings of ICASSP . Hong Kong: IEEE Press, 2003:p 200-203

[4] Zhu Xuan, Chen Yining, Liu Jia, et al. A novel efficient decoding algorithm for CDHMM-based speech recognizer on chip [C]. Proceedings of ICASSP . Hong Kong: IEEE Press, 2003:p 293-296

[5] Park, Junho (ISPL, Dept. of Electronics and Comp. Eng., Korea University); Ko, Hanseok. Effective acoustic model clustering via decision-tree with supervised learning[J] Speech Communication, v 46, n 1, Press May, 2005,:p 1-13

[6] Wang Guoliang, Liang,Weiqian, Liu Jia, et al. Moderate vocabulary English speech recognition system embedded on a chip [J]. Qinghua Daxue Xuebao/Journal of Tsinghua University. v 45, n 10, Press October, 2005:p 1393-1396

[7]Valtcho Valtchev. Discriminative Methods in HMM-based Speech Recognition [D]. St. John's College, University of Cambridge, 1995.

作者简介:姚竞(1982.4-),男,汉族,上海人,清华大学电子工程系硕士研究生, 研究方向: 语音信号处理与语音识别; 刘加(1954-),男,汉族,北京人,清华大学电子工程系教授,博士生导师,研究方向:语音识别/合成编码、语音识别专用芯片设计以及多媒体数字信号通信系统。

Biography:YAO Jing(1982.4-) male, shanghai, postgraduate in E.E Department of Tsinghua University,Research Aspect: Speech signal processing and speech recognition;

(100084 北京清华大学电子工程系)姚竞 王国梁 刘加

通讯地址:(100084 北京市海淀区清华大学 26#505 室) 姚 竞

(收稿日期:2008.4.05)(修稿日期:2008.5.22)