文章编号: 1003－0077(2004)06－0073－06

嵌入式语音识别系统的研究和实现①

方 敏, 浦剑涛, 李成荣, 台宪青

(中国科学院 自动化研究所高技术创新中心, 北京 100080)

摘要: 本文首先给出了一种适合于在嵌入式平台上实现的可变命令集的非特定人语音识别系统, 同传统的基于PC的非特定人语音识别系统相比, 该系统具备内存消耗小, 运算速度快的优点。然后给出了该语音识别系统在多种嵌入式平台上的实现和评估结果, 论证了非特定人语音识别系统在嵌入式平台上实现的可行性及其对硬件的最低配置要求, 在技术层次上分析了目前实现高性能语音识别SOC的主要问题和困难, 并指出了今后相关的研究方向。

关键词: 计算机应用; 中文信息处理; 嵌入式平台; 非特定人语音识别; 语音识别SOC

中图分类号: TP391.4 文献标识码: A

Research and Realization of Embedded Speech Recognition System

FANG Min, PU Jian-tao, LI Cheng-rong, TAI Xian-qing

(Hi-tech Innovation Center, Institute of Automation, Chinese Academy of Science Beijing 100080, China)

Abstract: Proposed in this paper is a novel speaker-independent speech recognition system, which is command-variable and suitable for realization based on embedded platform. Compared with traditional speaker-independent speech recognition system based on PC, our system is featured small storage and computation cost. The system is evaluated on several embedded platforms that are specially designed. According to the result of the evaluation, the feasibility of speaker－independent speech recognition system based on embedded platform is proved and the least requirement for the hardware is given. Then we analyzed the main problems and difficulties in the development of high performance speech recognition SOC (System On a Chip) from the point of technology, and pointed out some future works.

Key words: computer application; Chinese information processing; embedded platform; speaker-independent speech recognition; speech recognition SOC

1 前言

随着计算机软硬件技术、半导体技术、电子技术、通讯技术和网络技术等的飞速发展, 人类已经进入后PC时代。这个时代一个典型的特征就是: 各种新型智能化的设备日益广泛地走进人们的工作和生活, 而人与这些智能化终端之间的自然快捷稳定可靠的交互方式有助于提高人机交互的效率, 增强人对智能化设备的控制。作为人机交互最自然的方式, 语音技术的研究近几十年来取得了长足的进展, 其中语音识别由于其重要性和研究的难度更成为研究的热点[1,8]。

嵌入式语音识别系统是指应用各种先进的微处理器在板级或是芯片级用软件或硬件实现

① 收稿日期: 2003－08－01
基金项目: 863计划重点资助项目(2002AA118020); 北京市自然科学基金资助项目(4022010)
作者简介: 方敏(1980—), 男, 硕士研究生, 研究方向为嵌入式语音识别技术.

语音识别技术。语音识别系统的嵌入式实现要求算法在保证识别效果的前提下尽可能优化，以适应嵌入式平台存储资源少、实时性要求高的特点。实验室中高性能的大词汇量连续语音识别系统代表当今语音识别技术的先进水平。但由于嵌入式平台在资源和速度方面的限制，其嵌入式实现尚不成熟。而中小词汇量的命令词语音识别系统由于算法相对简单，对资源的需求较小，且系统识别率和鲁棒性较高，能满足大多数应用的要求，因而成为嵌入式应用的主要着眼点。

目前，在嵌入式平台实现了的主要是对系统的运算资源和存储资源要求比较低的特定人孤立词语音识别系统[7]。而在现实中，更多的语音识别应用要求系统具有非特定人的特点。相对而言，特定人语音识别系统可以对整词声学建模，识别则采用简单的 DTW 等匹配算法，这对小词汇量识别系统的实现效果比较理想。其缺点是，如果词表更换，就要求采集大量数据，重新训练模型，且训练好的模型又具有特定人的局限。本文介绍的非特定人语音识别系统采用基于汉语声韵母的声学建模单元，命令集可变，更换词表时无需重新训练模型，避免了特定人识别系统词表增大模型空间线性增加的缺点。

此项研究的目的在于：通过比较不同平台上的系统实现，分析语音识别系统嵌入式实现的最低运算和存储资源配置要求及系统优化方向，为语音识别系统板级及芯片级的设计开发提供参考依据。

为使系统尽可能少的占用嵌入式平台存储和运算资源并保证识别效果，我们对系统进行了优化，采用压缩的声学模型。我们分别在数字信号处理能力强的 DSP 平台和通用性好、性价比高的 ARM 平台上实现了该系统，考虑到不同的处理器及不同的硬件平台在系统时钟频率、数据处理速度、存储资源、缓冲机制等方面的差异，针对不同的平台对系统进行了模型大小及代码等方面的优化，进一步给出了在各平台上系统实现所要求的最小硬件资源配置及系统能达到的最高实时性能。嵌入式板级平台的测试评估结果为今后语音识别片上系统(SOC)的研制奠定了技术基础。

本文各小节内容安排如下：第二部分给出了一种适合于嵌入式平台实现的非特定人语音识别系统及其改进系统，第三部分分别介绍了三种嵌入式平台，第四部分给出了该非特定人语音识别系统的实验结果及其在不同嵌入式平台上的评估结果，并对结果进行了分析，最后是本项研究的阶段性结论，并对今后嵌入式语音识别技术的研究方向进行了探讨和展望。

2 适用于嵌入式平台的基于汉语声韵母建模的非特定人语音识别系统

汉语大词汇量连续语音识别系统(LVCSR)[2~6]一般采用以声韵母为建模单元的上下文相关的声学模型，一遍或多遍的搜索算法，以及 N-GRAM 的语言模型，词汇量一般达到几万个词，因此对运行平台的计算能力和存储能力要求非常高，目前只能在主流的 PC 机上运行。听写机曾是 LVCSR 的主流应用模式，但在实际应用中，由于语音识别引擎的识别率及其鲁棒性还不能达到应用的要求，因此听写机的应用并没有得到推广。但是，LVCSR 系统的与说话人无关和自然语言交互的特点，却始终是语音交互接口所不懈追求的。由于运算资源和存储资源，以及语音识别引擎本身性能的限制，目前要在嵌入式平台上实现一个可用的口语交互接口是很困难的。所以本文的研究集中在说话人无关上，希望能够在嵌入式平台上实现一个非特定人的语音识别系统。

2.1 BASELINE 系统(简称系统 1)

图 1 给出了非特定人语音识别 BASELINE 系统的框架结构。

首先，本系统的 BASELINE 可以看作是 LVCSR 的一个简化版本。具体简化是：忽略词间扩展，这样系统就成为一个命令词的语音识别系统；忽略语言模型，因为没有了词间扩展，语音识别引擎不再是连续的，语言模型也就不需要了；降低词汇量，因为一般而言，词汇量越小，词表的混淆度越低，识别引擎的识别率就越高，同时数据存储空间、搜索空间和计算量也就越小；采用不带音调的上下文无关声学模型，因为对于小词汇量而言，上下文无关的 BASEPHONE 模型在数据存储空间和计算量方面都要比上下文相关的 TRIPHONE 模型小得多，同时识别率也能够满足实际应用的要求，而采用音调会使模型的大小增加到原来的 5 倍，并且对口音敏感，因此也被忽略；把采样率从 16KHz 降为 8KHz，实验表明，对中小词表而言，采样率的降低所造成的识别引擎识别率的降低不超过 1%，但可以节省语音识别前端 50%的动态存储空间，减少运行时识别前端 25%的计算量。关于声学特征的选择，根据文献[7]中的实验结果，我们选择“能量＋MFCC＋一阶差分”，共 26 维，同 39 维的声学特征相比，节省了 1/3 的特征缓冲区空间。表 1 给出了该 BASELINE 系统的识别率测试结果。我们在基于 TI 公司的 TMS320C5409 DSP 的嵌入式平台上实现和评估了该系统，评估结果参考表 2。

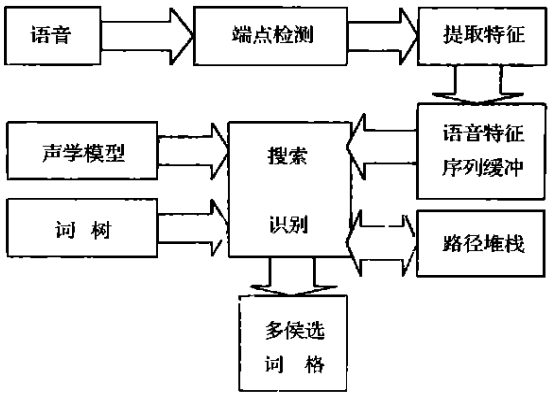

图 1 非特定人语音识别系统框架

2.2 改进后的系统（简称系统 **2**）

由 BASELINE 系统在 TMS320C540 平台上的评估结果可以看出，该系统对硬件平台的计算能力和存储能力的要求仍然很高。主要问题是，即便采用 BASEPHONE 模型，但声学模型仍然占用了系统整体存储空间消耗的 80%，声学得分的计算占用了搜索时间消耗的 90%，因此有必要对系统进行改进。这里研究了一种声学模型压缩算法，在几乎不降低系统识别率的前提下，对声学模型进行压缩，同时通过减少模型参数，降低声学得分运算的计算量。另外，通过优化某些数据结构，删除其中一些不必要的信息，能够节省大约 50%的动态空间。表 1 给出了模型压缩后不同压缩比下的系统识别率，可从中选择一种既能大大减小模型空间且对识别率影响不大的压缩比。这样就得到改进后的系统（系统 2）。表 2 给出了系统 2 在基于 TMS320C5402 DSP 的嵌入式平台上的评估结果。

CPU 采用不同的体系结构和指令集时，代码大小和执行效率都会相差很大。为了保证算法评估结果的可靠性，我们在当前比较流行的嵌入式中央处理器 ARM 平台上实现了上述优化后的非特定人语音识别系统。由于 ARM 处理器的对某些数学运算（如 LOG 函数）的处理能力远不如 DSP 强，大大影响了识别引擎的运行效率，因此对一些数学运算的函数进行了优化。表 2 给出了 ARM 平台的评估结果。

3 三种嵌入式平台描述

3.1 平台的硬件框架描述

平台的硬件结构如图 2 所示。

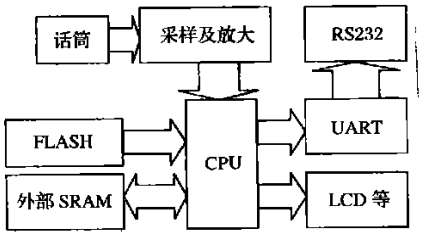

图 2　评估平台的硬性结构图

该平台包含:

a)一个 CPU 芯片

b)一片 FLASH

c)一个 CODEC 语音输入输出接口

d)一片 AD/DA 芯片

e)一个麦克风

f)如果 CPU 芯片的片内 RAM 存储空间太小,还必须外扩 RAM

g)如果需要向外设输出识别结果,可以增加 UART 接口和 RS232 接口

3.2　基于 **TMS320C5409** 的 **DSP** 嵌入式平台(简称 **DSP5409** 平台)

TMS320C5409 DSP 是 TI 公司 TMS320C54X 系列的产品。TMS320C54X 系列的 DSP 是一种典型的高性能、低功耗、16 位定点 DSP,广泛应用在各种嵌入式应用场合。54X 系列的 DSP 的处理器速度快,片内资源丰富,完全能够满足非特定人语音识别系统的要求。我们选择的 TMS320C5409 DSP 的处理器速度最高可达 100MIPS;片内共 48K 字的存储空间,其中 DRAM 是 32K 字,ROM 是 16K 字;片内具有丰富的外设,如 PLL, McBSP, DMA, HIP 等,其中 McBSP0 我们用来和 AD \DA 连接,接收采集到的语音数据。

外扩的资源有:两片 1M 字节的 8 位 FLASH,支持 16 位 BOOTLOADER 模式;外扩 512K 字 SRAM,其中 256K 字映射在程序区,供程序以扩展寻址方式访问,另外 256K 映射在数据区的高 32K 字的空间,分成 8 页访问,页面切换由烧录在外部 CPLD 中的逻辑控制。AD \DA 芯片采用 TI 公司的高速模拟接口芯片 TLC320AD50,该芯片支持多种采样率,包括 16KHz 和 8KHz,支持 16 位精度的采样,动态范围为 91dB。

3.3　基于 **TI320C5402** 的 **DSP** 嵌入式评估平台(简称 **DSP5402** 平台)

TI320C5402 DSP 也是 TI 公司 TMS320C54X 系列的产品,同 TI320C5409 DSP 相比,主要差别是,片内的存储空间要小得多,其中 DARAM 是 16K 字,ROM 是 4K 字。由于片内 DARAM 是影响系统的功耗和成本的主要因素,因此 TI320C5402 DSP 的功耗比 5409 DSP 更低,成本也只有 5409 DSP 的 1/3。TI320C5402 DSP 的处理器的速度最高也可达 100MIPS。

外扩的资源有:一片 64K 字的 16 位 FLASH,支持 16 位 BOOTLOADER 模式;外扩 64K 字的 SARAM,其中高 48K 字的空间为程序空间和数据空间共享,低 16K 字的空间的使用方法视 DSP 的中央处理器的配置寄存器的 OVLY 位的状态而定,当 OVLY=0 时,SARAM 低 16K 字的空间映射到程序区的低 16K 字空间,当 OVLY=1 时,程序空间和数据空间的低 16K 字共享 DSP 的片内 16K 字的 DARAM,片外扩展 SARAM 的低 16K 字将无法访问。由于程序在片内运行的速度比程序在片外运行的速度快 6~9 倍,需要把语音识别系统中某些运算量大的代码放到片内运行,因此我们选择 OVLY=1,这样,如果不考虑 FLASH 和 ROM 的话,程序空间和数据空间一共可用的 RAM 空间是 64K 字,这要比上述 TI320C5409 评估平台的存储资源小得多。AD \DA 芯片采用 AIC11,该芯片支持多种采样率,包括 16KHz 和 8KHz,支持 16 位精度的采样。另外,该平台对基于 TI320C5409 DSP 的评估平台上一些不必要的外部扩展资源进行了精简。

3.4 基于 **S3C4510b** 的 **ARM** 的嵌入式平台(简称 **ARM** 平台)

ARM 处理器采用三星公司的 S3C4510b 芯片,该芯片的主要特点是:采用 32 位 ARM7TDMI 内核[9],主频 50MHz,采用 RISC 指令集,包含 8kb 的可编程片内 Cache/SRAM,主要的片内外设包括:两路 HDLC 通道,两路 UART 通道,2 个 32 位定时器,18 个 GPIO。选择该芯片的主要考

虑是：处理速度较快，基本满足我们识别算法的要求；价格相当便宜，这对于该系统的商业化应用极具吸引力；功耗低。

其他硬件部件：外扩了 1 片 512K 的 SRAM，提供程序运行所需的临时空间；1 片 2M 的 FLASH(SST39VF160)存放程序代码及模型等数据；一个 CODEC 语音输入输出接口(16 位 ADC/DAC)；1 片 TI 公司的 TLC320AD50，用于采集语音数据。

4 实验和评估结果分析

4.1 压缩模型性能测试

首先以系统 1 为 BASELINE，我们测试了声学模型压缩对系统识别率的影响。测试环境描述如下：词表大小为 298 词，词长为 2～6 个字，平均为 3 个字；测试集采用实验室采集的孤立词测试集，共 2960 个孤立词，由 24 个说话人(14 男，10 女)采集得到。测试结果如表 1 所示，其中 BASELINE 采用未经压缩的模型(大小为 325KB)，系统识别率为 85.98%。压缩比为原模型与压缩后模型的大小比。

表 1 压缩模型性能测试结果

模型压缩比	模型大小(KB)	识别率
1∶1	325	85.98%
7.8∶1	41.47	85.67%
11.4∶1	28.47	85.79%
14.8∶1	21.97	85.06%
16.4∶1	19.87	85.67%
17.4∶1	18.72	84.42%

测试结果表明：该声学模型压缩算法能够在压缩比达到 11∶1 的情况下，基本保持系统在采用 CDHMM 模型的识别率。当模型压缩的更小时会对系统识别率有较大影响。因此在系统 2 中选用压缩比为 11.4∶1 的模型。

4.2 非特定人语音识别系统在嵌入式平台上的评估结果

需要说明的是，用于算法评估的嵌入式平台都是针对语音识别算法设计的，算法的改进，总是用资源更有限的硬件平台来实际验证之。系统 1 对资源的要求比较高，因此我们选择和设计了 TMS320C5409 DSP 嵌入式平台。系统 2 对资源的要求比系统 1 小得多，因此我们设计了基于 TMS320C5402 DSP 的嵌入式平台来实现和评估。TMS320C5402 和 TMS320C5409 DSP 的处理器速度可以根据需要在 10～100MIPS 之间选择。通过测试系统在不同处理器速度下的运行情况，发现当处理器速度降低到 30MIPS 时，语音识别前端仍能实时运行，而搜索引擎的运行为 1.3 倍实时，在可接受的范围内，因此又选用了主频为 50MHz(相当于 45MIPS)的基于 ARM 内核的 S3C4510B ARM 嵌入式平台，以验证评估结果，同时作为算法进一步优化的平台。表 2 给出了三次评估的最终结果。

表 2 系统在嵌入式平台上的评估结果

评估指标 \ 评估平台	DSP5409	DSP5402	ARM
前端耗时	0.31	0.31	1.5
搜索耗时	1.6	0.30	1.2
消耗片上 RAM	64KB	32KB	无
消耗片外 RAM	339KB	64KB	192KB
消耗 FLASH	365KB	64KB	167KB
CACHE	无	无	8KB

4.3 实验结果分析

1)系统 2 相对于系统 1 在速度和资源消耗方面的优势说明，系统的改进和优化是合理有效的；在资源有限的嵌入式平台上，完全有可能实现高性能的非特定人语音识别系统，这为将来在嵌入式平台上实现更为复杂的语音识别技术，如关键词检测等，奠定了基础。

2)从 ARM 平台的评估结果中，发现如下两个问题，一是 ARM 的处理速度比预想的要慢一倍左右，这说明，为语音识别系统选择 CPU 的时候，处理器的 MIPS 指标不能成为衡量其数据

处理速度的唯一指标；二是开发环境为 ARM 生成的可执行代码为 109KB，而同样的代码在 TMS320C54X 的开发环境下生成的可执行代码仅 22KB，是前者的 1/5，经分析，认为这是由于 ARM 采用了精简指令集的体系结构的缘故。

5 结论和展望

嵌入式语音识别系统具有广阔的市场应用前景。本文介绍的非特定人语音识别系统，相对于特定人孤立词语音识别系统具有多方面的优点，因此成为嵌入式语音识别系统研究和实现的主要着眼点。该系统的 BASELINE 是在 LVCSR 的基础上简化的，采用未压缩的模型，并在 TMS320C5409DSP 平台实现。为了使系统更适合于嵌入式应用，对 BASELINE 进行了模型的压缩和数据结构的优化，并在 TMS320C5402DSP 平台实现。系统改进后在 ARM 平台实现，也能基本满足实时性要求，且成本下降很多。通过这几种平台的系统测试，发现对 BASELINE 系统进行模型的压缩、数据结构的精简和代码优化之后，能大大降低系统实现平台的资源配置要求。同时，根据不同平台的自身特点(如 DSP 平台具有较强的信号处理能力，ARM 平台具有缓冲机制等)，对代码进行必要的优化。此项研究对于语音识别嵌入式模块的开发，对于今后研制嵌入式语音识别 API 及语音识别片上系统(SOC)具有很好的参考意义。

目前系统已在各个平台上实现并进行了综合评估，今后进一步的工作是：在语音识别算法方面，为了增强系统的环境鲁棒性，需要研究计算量和存储空间消耗都比较少的噪声消除或补偿算法、可靠的集外词和噪声的拒识算法等；在嵌入式平台方面，研发语音识别前端的专用处理模块，使其能执行更为复杂的语音信号前端处理算法。

参 考 文 献：

[1] Lawrence Rabiner, Biing—Hwang Juang. 语音识别基本原理(影印版)[M]. 北京：清华大学出版社，1999.

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

[3] 高升. 语境相关的声学模型和搜索策略的研究[D]. 中国科学院图书馆：中国科学院自动化研究所博士学位论文，2001.

[4] 高升，徐波，黄泰翼. 基于决策树的汉语三音子模型[J]. 声学学报，2000，25(6).

[5] 马龙. 汉语命令词识别，关键词检测的研究与应用[D]. 中国科学院图书馆：中国科学院自动化研究所硕士学位论文，2002.

[6] 易克初，田斌，付强. 语音信号处理[M]. 北京：国防工业出版社，2000.

[7] 丁国宏，李成荣，徐波. 非特定人孤立词语音识别系统在定点 DSP 上的应用[A]. 第六届全国人机语音通讯会议[C]，2001.

[8] B. H. Juang. The past, present and future of speech processing[J]. IEEE Signal Processing Magzine, May, 1998.

[9] [英] Steve Furber 著，田泽等译. ARM Soc 体系结构[M]. 北京：北京航空航天大学出版社，2002.