基于DHMM的嵌入式语音识别系统的实现与优化

崔金钟[1]，周远彬[1]，陈雷霆[1,2]

(1. 电子科技大学计算机科学与工程学院　成都　611731;　2. 东莞-电子科技大学电子信息工程研究院　广东 东莞　523808)

【摘要】为满足在嵌入式系统上实现非特定人的语音识别需求，介绍一种采用离散隐马尔科夫模型(DHMM)的嵌入式语音识别系统的实现方法。设计出一种在嵌入式系统上完成语音识别功能、在PC系统上完成模板训练功能的软件架构。针对TMS320VC5509A处理器的特点，对识别软件进行实现和优化。根据语音信号不可逆的特点，提出一种优化的viterbi算法。实验结果表明，在保证识别率的条件下系统取得了200 ms以内的实时响应速度，具有很高的实用性。

关　键　词　离散隐马尔科夫模型;　语音识别;　TMS320VC5509A;　viterbi算法

中图分类号　TN912　　文献标志码　A　　doi:10.3969/j.issn.1001-0548.2013.06.023

Implementation and Optimization of Embedded Speech Recognition System Based on DHMM

CUI Jin-zhong[1], ZHOU Yuan-bin[1], and CHEN Lei-ting[1,2]

(1. School of Computer Science & Engineering, University of Electronic Science & Technology of China　Chengdu　611731;

2. Institute of Electronic and Information Engineering in Dongguan, University of Electronic Science & Technology of China　Dongguan Guangdong　523808)

Abstract　In order to meet the demand of realizing the speaker-independent speech recognition on the embedded system, a embedded speech recognition system in discrete hidden Markov model (DHMM) is proposed. The software architecture of the recognition software part running on embedded system and the training software part running on PC system are designed. According to features of TMS320VC5509A processor, the recognition software on embedded system is implemented and optimized. Also it is presented that the Viterbi algorithm is optimized on the basis of the irreversible characteristics of the speech signal. Experiments show the real-time response speed of the system achieved within 200 ms at same the recognition rate, and show its high availability.

Key words　DHMM;　speech recognition;　TMS320VC5509A;　viterbi algorithm

随着嵌入式系统的广泛应用以及智能人机交互技术的发展，语音识别技术在嵌入式系统上的需求逐步增加，因此近年来语音识别技术成为一个研究热点。在资源丰富的PC机上语音识别系统的识别率和识别时间都基本满足了实际需求，但在资源有限的嵌入式设备上如何提高非特定人语音识别率和减少识别时间却成为研究的难点和重点。

目前典型的语音识别算法包括：动态时间规整(DTW)算法、隐马尔科夫模型(HMM)、人工神经网络(ANN)。文献[1-3]研究了在嵌入式处理器上实现与优化DTW算法，但基于DTW算法的语音识别有其局限性，只能实现特定人简单孤立词的识别。文献[4-5]研究了基于连续隐马尔科夫模型(CHMM)的嵌入式语音识别系统，由于CHMM采用了高斯分布函数，计算量较大，难以在单独嵌入式处理器系统上达到实时性，因此文献[4-5]采取了硬件加速设计、FPGA等技术来解决这个问题，使得系统较为复杂，硬件成本高。文献[6-7]研究了SVM在语音识别中的应用，这些方法同样存在语音识别时间长的问题，难以在嵌入式系统上保证其实时性。

本文实现的嵌入式语音识别系统，通过采用离散隐马尔科夫模型(DHMM)解决了DTW算法只能实现特定人、小词汇量的识别问题，并且通过采用识别部分和模板训练部分进行分离的方式，设计出一种在嵌入式系统上完成语音识别功能、在PC系统上完成模板训练功能的软件架构，简化了嵌入式系统的复杂性；同时针对16位定点DSP处理器的特点，重点研究了在嵌入式部分上语音识别算法的优化，通过优化viterbi算法与优化存储空间等技术手段，解决了系统的实时性指标。这样的嵌入式语音识别系

收稿日期：2012 – 02 – 21；修回日期：2013 – 09 – 16
基金项目：中国教育科研网格(ChinaGrid)项目
作者简介：崔金钟(1966 –)，男，教授，主要从事嵌入式计算机系统方面的研究.

统设计方案既达到了识别率和识别时间的要求，又使系统简化，降低了硬件成本。

1　系统硬件设计

本文系统的硬件结构框图如图1所示。核心芯片为通用16位定点DSP芯片：TMS320VC5509A(以下简称VC5509A)属于TI公司TMS320C55x系列。VC5509A最高工作频率为200 MHz，包括3条内部读总线、2条内部写总线。TMS320VC5509A存储空间包括片上储存空间：32K×16位单周期双端口访问DARAM、96K×16位单周期单端口访问SARAM、32K×16位ROM，以及扩展存储空间。扩展存储空间通过EMIF接口完成，系统扩展4M ×16位SDRAM。VC5509A具有丰富的片上外设：6个DMA通道、3个McBSPs通道、AC97接口、ADC接口、USB2.0从接口、CPLD扩展接口等。

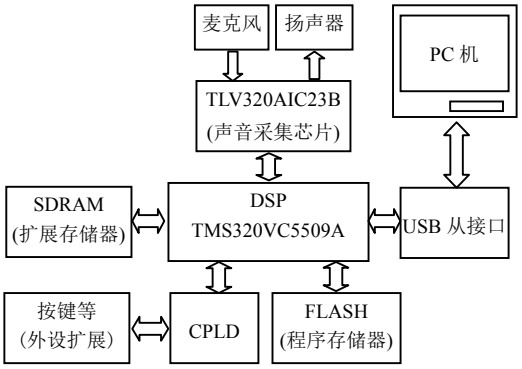

图1　系统硬件结构图

系统采用TI推出的一款高性能的立体声音频Codec芯片TLV320AIC23B(以下简称AIC23B)。内置耳机输出放大器，支持麦克风和立体声两种输入方式，并且可编程增益调节。VC5509A通过IIC总线接口与AIC23B接口连接，实现对其寄存器设置。VC5509A的McBSP0配置为IIS工作方式连接AIC23B的数据接口，实现数据交换。

语音识别系统包括模板训练和语音识别两大部分。本文设计为了减轻嵌入式系统的运算量，首先将语音识别系统分为训练和识别部分，只将必需的识别部分放在嵌入式系统上运行，而将对存储空间需求大，计算量较大的训练部分放置在PC机上进行完成。PC系统训练完成后的模板数据通过USB接口下载到嵌入式系统，并烧写到FLASH芯片中。嵌入式系统启动后，训练数据被加载到内存中，供识别使用。

2　系统软件设计

为了避免连续型隐马尔科夫模型中高斯函数的巨大计算量，同时根据系统非特定人孤立词语音识别的特点，本文设计采用了离散隐马尔科夫模型。

2.1　离散隐马尔科夫模型

一个离散隐马尔科夫模型[8]可以用以下参数表示：1) N为模型的状态个数。状态的集合$S=\{S_1, S_2,\cdots,S_N\}$，$t$时刻的状态表示为$q_t$。2) M为观察符号个数。即每个状态可能输出的观察符号数目。观察符号集合表示为：$O=\{O_1,O_2,\cdots,O_T\}$。3) A为状态转移概率分布：

$$A=\{a_{ij}\}, a_{ij}=P[q_{t+1}=S_j \mid q_t=S_i], 1\leqslant i, j\leqslant N \tag{1}$$

$$a_{ij}\geqslant 0 \text{且} \sum_{j=1}^{N} a_{ij}=1 \tag{2}$$

4) B为状态S_j的观察符号概率分布：$B=\{b_j(O_k)\}$，$b_j(O_k)$表示在t时刻处于状态S_j输出概率为O_k的概率。5) π为初始状态分布：$\pi=\{\pi_i\}$，$\pi_i=p[q_1=S_i]$，$1\leqslant i\leqslant N$。DHMM常用$\lambda=\{A,B,\pi\}$来简记。

2.2　系统软件结构设计

系统的软件结构如图2所示。图2的软件结构图包括嵌入式系统和PC系统，两部分通过USB接口连接。PC系统完成码本训练和语音模板的训练，嵌入式系统只完成语音识别部分。

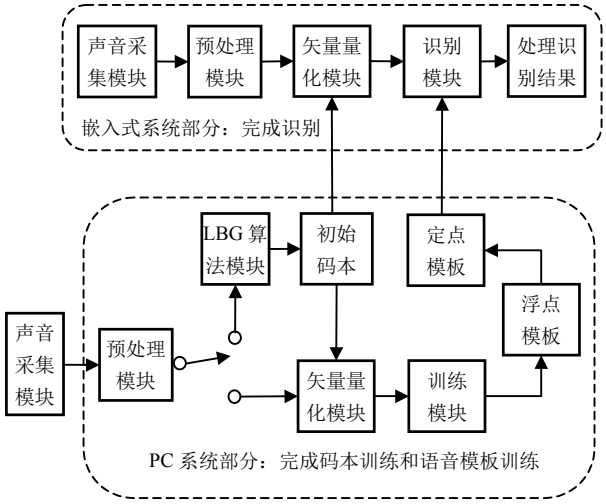

图2　系统软件结构图

系统形成码本处理流程：

1) 采集语音样本。

2) 语音经过预处理模块进行端点检测，然后提取语音的Mel频率倒谱系数(mel frequency cepstrum coefficient，MFCC)[8]。

3) 用LBG[8]算法模块训练出码本。

系统训练语音模板数据流程：

1) 采集词汇语音样本。

2) 将样本语音用预处理模块提取语音MFCC特征参数。

3) 将步骤2)提取的语音MFCC参数码本矢量量化[8](VQ)。

4) 语音特征参数矢量量化后，通过训练模块训练出语音模板，定点化后下载到嵌入式系统中。

嵌入式系统上完成系统的识别的流程为：

1) 采集待识别词汇的语音数据。

2) 提取语音MFCC特征参数。

3) 将提取的语音MFCC参数矢量量化，形成观察序列。

4) 用识别模块计算识别结果。

由于VC5509A是一块16位定点DSP处理器，处理浮点算法将会调用大量的浮点模拟库，系统难以达到实时性要求，因此本文设计对于嵌入式系统的软件采用算法定点优化。

3 viterbi算法的优化

如图2所示的系统软件结构图中，嵌入式系统识别时间主要花在预处理模块和识别模块两部分。预处理模块采用算法定点化得以解决。然而经过实验统计分析，系统识别一个词，约60%左右的CPU时间用在识别模块的viterbi算法上，对viterbi算法的改进和优化对提高系统实时性起着至关重要的作用。

3.1 viterbi算法

Viterbi 算法解决给定一个观测值序列 $O=o_1,o_2,\cdots,o_T$ 和一个模型 $M=\{A,B,\pi\}$，确定最佳的一个状态转移序列 $S=s_1,s_2,\cdots,s_T$ 的问题。算法如下[8]：

1) 初始化：

$$\delta_1(i)=\pi_i b_i(O_1) \qquad 1\leqslant i\leqslant N,\psi_1(i)=0 \tag{3}$$

2) 迭代计算：

$$\delta_t(j)=\max_{1\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}]b_j(O_t) \quad 2\leqslant t\leqslant T,1\leqslant j\leqslant N \tag{4}$$

$$\psi_t(j)=\arg\max_{1\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}] \quad 2\leqslant t\leqslant T,1\leqslant j\leqslant N \tag{5}$$

3) 最后计算：

$$p^*=\max_{1\leqslant i\leqslant N}[\delta_T(i)] \tag{6}$$

$$q_T^*=\arg\max_{1\leqslant i\leqslant N}[\delta_T(i)] \tag{7}$$

4) 路径回溯：

$$q_t^*=\psi_{t+1}(q_{t+1}^*),t=T-1,T-2,\cdots,1 \tag{8}$$

算法中T为语音帧数；N为DHMM模型的状态个数(系统中选取的状态数N为6)。算法中步骤2)的式(4)、式(5)的迭代需要完成$T\times N\times N$次，因此只要减少其迭代次数，就可以提高系统识别的实时性。

3.2 viterbi算法优化

文献[9]根据viterbi算法中的状态转移矩阵$\boldsymbol{\psi}$，研究了在不改变标准viterbi算法运行时间的条件下减少内存需求的解决方法。本文对viterbi算法改进的主要目的则是减少算法运行时间。

3.1节分析viterbi算法中最耗时部分是算法步骤2)，如果要减少其计算量，关键是减少$T\times N\times N$次的迭代次数。Viterbi算法解决一个观察序列匹配模型的一个最佳状态序列，而这个最佳状态序列用一个状态转移矩阵$\boldsymbol{\psi}$表示。一个典型的状态转移矩阵$\boldsymbol{\psi}$如图3所示。

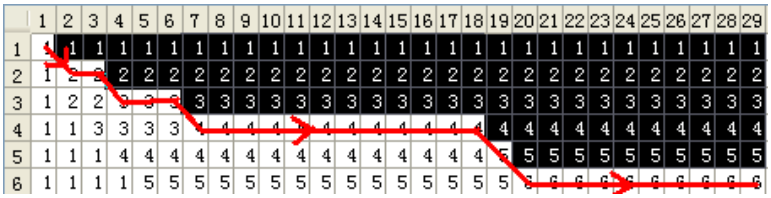

图3 典型的状态转移矩阵图

图3中横坐标代表语音帧序号，纵坐标代表状态数。图3中有6个状态，29帧语音数据。系统的状态转移序列为：1、2、2、3、3、3、4、4、4、4、4、4、4、4、4、4、4、5、6、6、6、6、6、6、6、6、6(图3中带箭头的折线所示)。图中带箭头折线的走向表示观察序列沿着这条路径的输出概率最大，这就是viterbi算法计算出的最佳状态序列。

确定最佳状态序列的原理是从时刻t到时刻t+1选择的路径是使得输出概率最大的路径。考虑系统选取的状态转移为从左到右不可逆型，如图4所示。

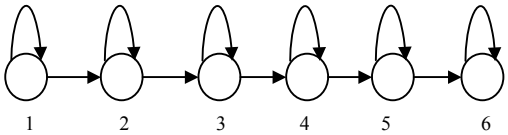

图4 系统的状态转移图

图4中，系统选取从左到右型DHMM。根据这个特点可以确定：系统的状态转移走向只能向当前状态或者下一个状态转移。由于语音信号的不可逆性，所以当前状态不可能转移到前续状态，这样图3

中带箭头的折线表示最佳状态序列曲线的走向，不可能向图中的黑色区域转移。而经典viterbi算法对所有的情况都计算在内，存在大量冗余的计算。改进后的viterbi算法大大减少了计算量，算法如下：

1) 初始化：

$$\delta_1(i)=\pi_i b_i(O_1) \qquad 1\leqslant i\leqslant N, \psi_1(i)=0 \tag{9}$$

$$\mathrm{st}=1 \tag{10}$$

2) 迭代计算：

$$\delta_t(j)=\max_{\mathrm{st}\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}]b_j(O_t) \quad 2\leqslant t\leqslant T, \mathrm{st}\leqslant j\leqslant N \tag{11}$$

$$\psi_t(j)=\arg\max_{\mathrm{st}\leqslant i\leqslant N}[\delta_{t-1}(i)a_{ij}] \quad 2\leqslant t\leqslant T, \mathrm{st}\leqslant j\leqslant N \tag{12}$$

$$\mathrm{st}=(\psi_t(i)==i) \qquad 1\leqslant i\leqslant N \tag{13}$$

3) 最后计算：

$$p^*=\max_{1\leqslant i\leqslant N}[\delta_T(i)] \tag{14}$$

$$q_T^*=\arg\max_{1\leqslant i\leqslant N}[\delta_T(i)] \tag{15}$$

4) 路径回溯：

$$q_t^*=\psi_{t+1}(q_{t+1}^*) \qquad t=T-1,T-2,\cdots,1 \tag{16}$$

算法增加一个st变量，记录当前需要参与计算的开始状态序号。在序号为st以前的状态都不需参与迭代计算，这样匹配一个模型可以减少60%以上的计算量。

4　存储空间优化

TMS320VC5509A芯片的存储空间组织如图5所示。包括3类不同的存储器：DARAM、SARAM、SDRAM，其访问速度依次减慢而容量却依次增大。

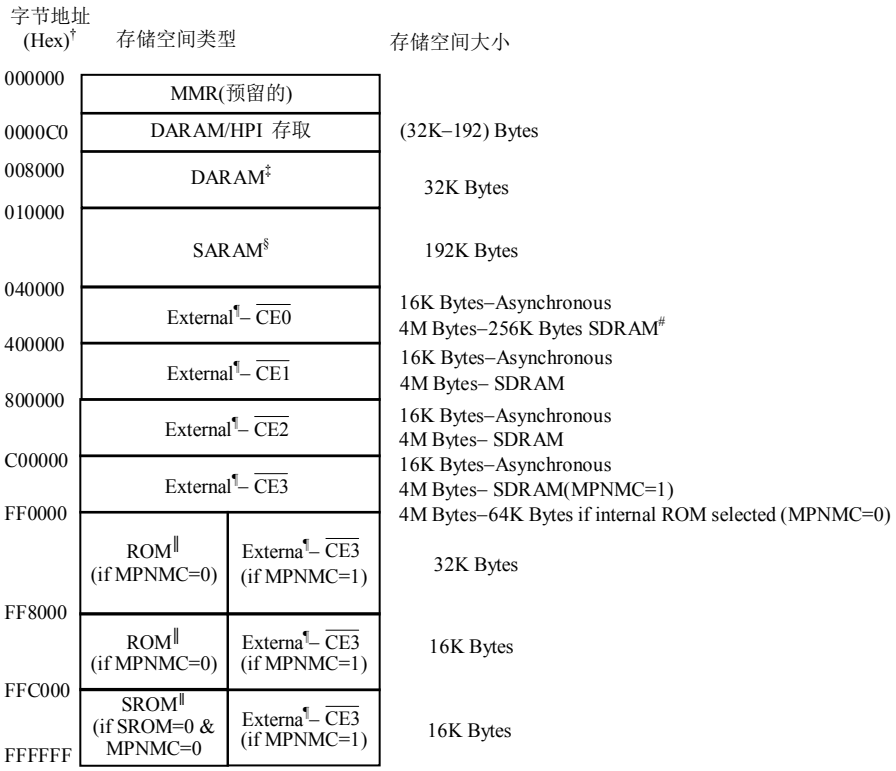

图5　VC5509A存储空间组织

DARAM提供单周期双端口访问、SARAM提供单周期单端口访问。为了提高系统的实时性指标，尽可能利用高速存储器。

软件代码经过系统编程工具CCStudio V3.3编译后形成的执行代码分为：text、cinit、data、bss、stack、sysstack、sysmem等段，这些段存储空间的安排，决定了系统的效率。

程序运行过程中的关键部分以及频繁使用部分，安排到存储器DARAM中。结合各段使用频率与编译后各段的大小，将stack、sysstack、bss、switch、const等段的空间分配到DARAM中；将text段、cinit段、data段、fftcode段、模板数据、MFCC空间、viterbi动态空间、采样语音数据空间分配到SARAM中；sysmem段分配到SDRAM中。详细的存储空间组织图如图6所示。

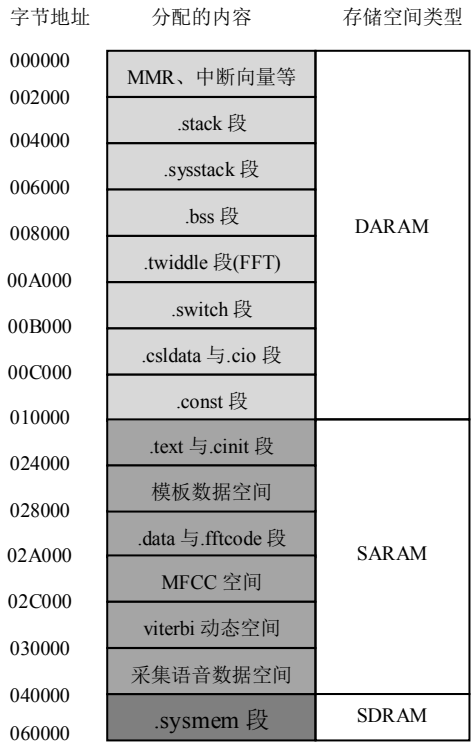

图6 系统存储空间组织

5 实验结果

实验采用100个孤立词汇作为语音识别词汇，记录从语音采样成功到识别出结果系统的运行时间，对系统采用未优化viterbi算法与优化viterbi算法进行比较。实验结果如图7所示。

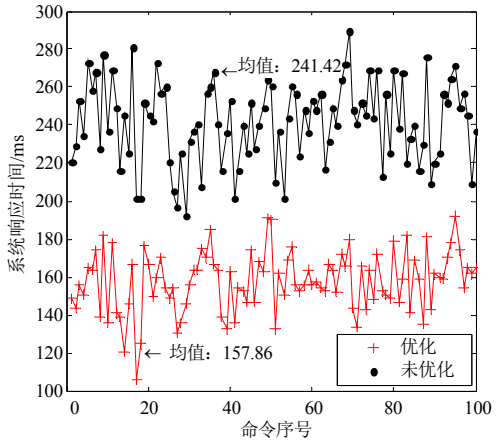

图7 viterbi算法优化与未优化的识别时间比较

图中，统计100个孤立词汇在系统优化和未优化的响应时间得出以下结论：系统优化后的响应时间较未优化有明显的提高，优化后的系统实时性平均比未优化的系统实时性提高83.56 ms，使整个系统的响应时间提高约30%。因此系统采用优化后的viterbi算法，实时性得到显著增强，而系统的识别率均保持在92%左右，完全可以满足实际需求。

6 结 论

本文采用DHMM为基础，构建了一个非特定人孤立词语音识别系统。针对孤立词语音识别与16位定点DSP芯片的特点，设计的语音识别系统由嵌入式系统与PC系统组成。嵌入式系统只完成语音识别功能，而把语音模板训练在PC系统上完成。本文着重解决嵌入式系统的实时性。根据语音信号的不可逆的特点，对viterbi算法进行改进以提高系统实时性，同时采用算法的定点化、存储空间的优化等技术手段，使系统在保证90%的识别率以上的条件下，可以在200 ms以内完成100个词汇量的语音识别，达到了很好的实用性

参 考 文 献

[1] LIU Bai-fen. Research and implementation of the speech recognition technology based on DSP[C]//AIMSEC, 2011 2nd International Conference. [S.l.]: [s.n.], 2011: 4188-4191.

[2] QU Qing-lin, LI Liang-guang. Realization of embedded speech recognition module based on STM32[C]//ISCIT, 2011 11th International Symposium. [S.l.]: [s.n.], 2011: 73-77.

[3] ZHANG Jing. Research of improved DTW algorithm in embedded speech recognition system[C]//ICICIP, 2010 International Conference. [S.l.]: [s.n.], 2010: 73-75.

[4] LI Peng, TANG Hua. Design of a low-power coprocessor for mid-size vocabulary speech recognition systems[J]. Circuits and Systems I, IEEE Transactions, 2011, 58(5): 961-970.

[5] CHENG O, ABDULLA W, SALCIC Z. Hardware-software codesign of automatic speech recognition system for embedded real-time applications[J]. Industrial Electronics, IEEE Transactions, 2011, 58(3): 850-859.

[6] QU Zhi-yi, LIU Yu, ZHANG Li-hong. A speech recognition system based on a hybrid HMM/SVM architecture[C]//Proceedings of the First International Conference on Innovative Computing, Information and Control(ICICIC). [S.l.]: [s.n.], 2006, 2: 100-104.

[7] SOLERA U R, GARCIA M A, PELAEZ M, et al. Real-time robust automatic speech recognition using compact support vector machines[C]//Audio, Speech, and Language Processing, IEEE Transactions. [S.l.]: [s.n.], 2011.

[8] HUANG Xue-dong, ALEX A, HON Hsiao-wuen. Spoken language processing: a guide to theory, algorithm and system development[M]. NewJersey: Prentice Hall PTR, 2001.

[9] BOBBIN J. An incremental viterbi algorithm[C]//ICMLA 2007, Sixth International Conference. [S.l.]: [s.n.], 2007: 104-111.

编 辑 漆 蓉