HMM非特定人孤立词语音识别系统的片上实现*

张晨燕[1]，孙成立[1,2]
（1.石家庄经济学院信息工程学院 石家庄 050031；2.北京邮电大学信息工程学院 北京 100876）

摘 要 在SEED-DEC5502 DSP嵌入式系统开发平台上实现了一个面向非特定人的孤立词语音识别系统，与传统的基于特定人的语音识别系统相比，该系统无需用户训练，易于使用。系统采用改进的基于语音对数域能量变化率的实时端点检测算法，仅对检测的有声段语音进行特征提取和解码，减少了要处理的语音帧数 对状态输出概率计算进行了分析和优化，进一步降低了计算负担。实验表明系统在100词条的情况下识别率达到98%，识别时间为1.03倍实时。

关键词 语音识别；嵌入式系统；端点检测；状态发射概率

1 引言

随着语音识别技术的发展，HMM（hidden markov model）在语音识别上得到了成功应用。目前孤立词语音识别技术已经趋于成熟，中、小词表（词表容量为10~100个）的识别率已经达到98%以上，孤立词语音识别技术已经由PC机走向嵌入式应用。孤立词识别根据词表模型的建立方式分为特定人和非特定人2种方法。特定人识别技术不用存储声学模型，占用空间少，但要求用户对词表进行训练后才能使用，而且词表一旦发生变化需要用户重新训练。采用这种方法的系统使用非常不方便，目前应用领域也只限于玩具和一些小型消费类电子产品。针对特定人识别方法的不足，许多科研机构相继开始研发基于非特定人的语音识别芯片，2000年美国TI公司开发出以TMS32054x系列DSP为核心的嵌入式非特定人语音识别芯片，该系统英文连续数字串的识别率为98.2%，34条英文控制指令的识别率为98.4%[1,2]，目前国内也一直在进行这方面的研究[3,4]。本着这个目的，笔者开发了基于TMS320VC5502 DSP嵌入式开发平台的非特定人孤立词语音识别系统，对100个词汇量的词表，系统的正确识别率能达到98%。该项技术不需要用户进行任何训练，而且系统的词表可以由用户自己定制和改变。无疑该技术更加方便用户操作，技术推广性强，很容易移植到手机等嵌入式通信产品中。

下面就非特定人孤立词识别系统的原理及其嵌入式系统实现分别叙述。

2 孤立词语音识别的基本原理

图1为孤立词语音识别的原理框图，其中预处理部分包括语音信号的模/数转换、分帧、加窗、端点检测等过程。特征提取部分是整个语音识别系统的基础，对语音识别率有极其重要的影响。常用的特征提取手段有LPC(linear prediction coefficient，线性预测特征参数)和MFCC(mel-frequency cepstrum coefficient，美尔域倒谱特征参数)。语音识别的过程可以被看作是模式匹配（或语音解码）的过程。模式匹配是指根据一定的相似性度量法则，使未知模式与参考模式库中的某一个参考模型获得最佳匹配的过程，对于基于HMM的系统都采用Viterbi算法来进行解码。

参考模式库记载词表发音的所有声学信息，对于基于非特定人的孤立词识别系统，参考模型通过基本声学模型拼接而成。

* 河北省科技厅资助项目（No.052135147），河北省科技厅指导性项目（No.042135105）

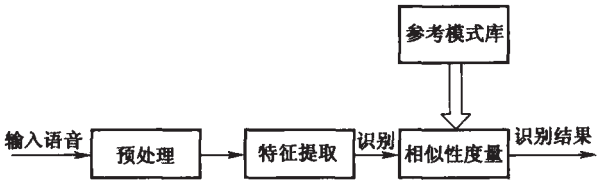

图 1 孤立词语音识别的原理

如“北京”发音的参考模型可以看作由 4 个声、韵母的 HMM 模型 b+ei+j+ing 拼接而成。

3 系统的硬件平台和采集方式

嵌入式平台采用 SEED-DEC5502 EVM 板，TMS320C5502 是一款高性能、低价位、低功耗（仅为 0.25 mW）的 16 位定点 DSP 芯片，C5502 采用变长指令，增强型并行机制。从结构上看，C5502 具有改进了的哈佛结构，拥有 2 个乘加单元（MAC），运行速度最高可达 400 MIPS。它有 64 Kbyte 的片内 DRAM，6 个 DMA 通道，1 个 I²C 接口，3 个 McBSP（multi-channel buffered serial port，多功能缓冲串行接口）分别为 McBSP0、McBSP1 和 McBSP2，适合中等词汇量的孤立词识别引擎对处理器资源的需求。C5502 程序/数据空间采用统一编址方式，采用 24 位地址线，寻址空间可达 16 Mbyte。

输入语音通过音频采集芯片 TLV320AIC23B 按照 8 000 Hz 采样频率、16 bit 量化单位采样格式采样，采集的数据通过 DSP 的 McBSP 连接至 DMA 控制器指向的数据缓冲区，该缓冲区可以为 DSP 的内存或外存。AIC23B 是 TI 推出的一款高性能的立体声音频 Codec 芯片，内置耳机输出放大器，支持 MIC 和 LINE IN 2 种输入方式（二选一），且对输入和输出都具有可编程增益调节。AIC23B 的模/数转换（ADC）和数/模转换（DAC）部件高度集成在芯片内部，采用了先进的 Sigma-delta 过采样技术，可以在 8~96 kHz 的频率范围内提供 16 bit、20 bit、24 bit 和 32 bit 的采样，ADC 和 DAC 的输出信噪比分别可以达到 90 dB 和 100 dB。与此同时，AIC23B 还具有很低的功耗，回放模式下功耗仅为 23 mW，省电模式下更是小于 15 μW。McBSP 是一种功能很强的同步串行接口，具有很强的可编程能力，可以直接配置成多种同步串口标准，与各种器件实现无缝接口，DMA 控制器通过同步接收事件接收 McBSP 的先入先出寄存器的每一个采样点数据。由于语音识别都是按帧对数据进行集中处理的，为了实现信号的实时处理，笔者把整个数据存储缓冲区大小设为 640 个字，同时令 DMA 接收中断控制器的半帧中断位和整帧中断位设置为 1，用来控制 DMA 的半缓冲区满中断和整个缓冲区满中断，这样，采集的语音数据每满半个缓冲区（320 字）将触发 DMA 中断程序，要求对这半个缓冲区的数据进行信号处理。与此同时，另一半缓冲区在不断地存储即时采集的语音信号，通过这样周而复始从而实现语音信号的批量采集- 处理的并行工作模式。

4 系统主要功能模块的实现

4.1 端点检测和特征提取

在语音识别系统中，数字语音信号是由语音、静音和各种背景噪音混合组成的。在这种信号中将语音和各种非语音信号时段区分开来，准确地确定出语音信号起始点和终点被称之为端点检测。在语音识别中，端点检测的性能对于识别正确率和识别速度都有着重要的影响。实时系统要求采用的 VAD（端点检测）算法也必须是实时的，为了提高系统的运行速度，识别程序仅对 VAD 算法检测出语音的活动区间段的语音进行识别。本系统采用的 VAD 算法是根据欧洲电信标准协会（ETSI）提供的 VAD 算法改进而来[5,6]，该算法能够根据每帧对数能量值的变化率来判断当前帧是否为活动语音帧。调用该 VAD 算法对每一帧语音检测的返回结果 VadFlag 为一个二值布尔变量，VadFlag=1 表示检测有语音，VadFlag=0 表示没有检测到语音。与传统的基于能量和过零率的方法相比，基于 LOG（对数）域能量变化的 VAD 算法顽健性比较好，抗噪声能力强，目前已经应用在电话信道的语音激活检测方面。

语音识别特征参数采用 MFCC 倒谱系数，MFCC 利用了人耳的感知特性，人耳对频率的区分能力大致和频率成对数关系。MFCC 特征主要反映语音的静态特征，语音信号的动态特征可以用静态特征的一阶差分谱和二阶差分谱来描述，这些动态信息和静态信息相互补充能很大程度地提高系统的识别性能。

由于 ETSI 提供的算法应用对象为电话信道，信噪比小，而嵌入式设备的应用环境的信噪比较大。实际中对原算法进行了改进，在检测语音起点时，只有当检测到连续 10 帧 VAD 返回结果为 1 时，才认为是检测到语音的起点，否则认为是噪声干扰的影响而不认为是活动语音。在端点检测的同时，系统对语音数据实时提取 13 维的 MFCC 静态特征。在检测到语音的起点后，如果检测到连续 20 个 0，就认为是检测到语音的终点，同时结束静态 MFCC 特征提取。接下来将提取 MFCC 动态特征，计算 MFCC 系数的一阶差分和二阶差分特征参数，并进行倒谱均值归一化来减少声道对特征的影响。最后得到的 MFCC 系数总共 39 维。实践证明改进的 VAD 算法的适应性较好，顽健性强，有效解决了原有算法 VAD 模块对噪声敏感的问题。

4.2 语音解码部分的优化

通过对程序进行分析发现，系统所花费的时间主要消耗在语音解码程序上[7]。而在计算 Viterbi 算法时，尤其以计算状态输

出概率的计算复杂度最高。

已知模型 λ的第 s 个状态 λ_s 的输出概率的计算公式如式(1)所示。

$$P(x_t|\lambda_s)=\sum_{m=1}^{M}C_m\cdot\exp\left[-\frac{1}{2}\sum_{j=1}^{J}\frac{((x_t(j)-\mu_m(j))^2}{\sigma_m^2(j)}\right] \quad (1)$$

式(1)中，μ_m、σ_m^2分别为 λ_s 状态下第 m 个独立的高斯混合概率密度的均值和方差，C_m为高斯加权系数。C_m对于每一个高斯混合密度而言是一个常量。由于每个词的声学模型由声韵母拼接而成，每个声韵母的声学模型由 3 个状态组成，每一个状态又含有 4 个相互独立的高斯型混合概率密度。在计算状态的似然概率过程中，要遍历词模板中的所有状态，并遍历每个状态所包含的所有高斯型混合概率密度，然后将各个高斯混合分量的似然概率乘以混合分量后相加作为整个状态的似然概率输出，这样就造成了搜索过程计算量过大，尤其是在计算状态似然概率的过程中对各个高斯混合分量的似然概率进行对数加法运算消耗了大量的运算资源[8]。经过大量实验分析发现，整个状态的似然概率与最大高斯混合分量产生的概率是非常接近的，也就是说，最大高斯混合分量在确定状态的似然概率过程中起到了决定性的左右，选取最大高斯混合分量法的基本思想就是在得到了各个高斯混合分量的似然概率以后，选取贡献最大的分量直接作为该状态的似然概率输出，从而省去了对数加法这一复杂的计算过程，节省了系统资源，提高了识别速度。另外一个值得注意的问题是：由于 DSP 没有专用的除法器，所以除法运算要比乘法运算所消耗时钟周期多几十倍，所以，在存储声学模型时，直接存储已经计算好的方差的倒数，以便将除法运算转化成乘法运算。经过简化后，在对数域下式(1)的简化计算方法可以表示为：

$$\ln(p(x_t|\lambda_s))\approx\max_{1\le m\le M}\left\{\ln(C_m)-\frac{1}{2}\sum_{j=1}^{J}(x_t(j)-\mu_m(j))^2\cdot\frac{1}{\sigma_m^2(j)}\right\} \quad (2)$$

5 试验结果及分析

实验用到的测试集词长范围为 2~5 字，平均字长为 2.8。实验中采用的声学模型为无音调的 monophone，共计 61 个音素，每个音素模型由 3 个实状态、2 个虚状态构成，而每个状态由 4 个混合分量构成。提取的 MFCC 系数为 39 维。

测试集分别选自 5 男 5 女的录音数据，共计 500 句。这些发音都是在实验室环境下通过 EVM 板的麦克风接口录音的，表1、表 2、表 3 是 C5502 时钟频率为 200 MHz(时钟周期 5 ns)时相关的试验结果比较。

表 1 系统的实时处理功能模块平均消耗时钟比较

	VAD	MFCC(静态)
程序消耗时钟数 (个)	1 215	12 000

表 2 基线系统的平均识别正确率和识别速度

词表容量(个)	识别正确率	识别速度(实时系数)
50	98.6%	1.22
100	98.2%	1.44
150	97%	1.98

表 3 取最大高斯混合分量的平均识别正确率和识别速度

词表容量(个)	识别正确率	识别速度(实时系数)
50	98.6%	0.89
100	98.0%	1.03
150	96.6%	1.23

从表 1 的数据可以看出，ETSI 所提供的 VAD 程序所花费时钟很少，而计算静态 MFCC 系数的程序所花时间<20 000 个时钟周期，也能满足实时计算要求。从表 2 和表 3 的数据比较发现，采用取最大高斯混合分量的方法在精度稍微下降的情况下，系统速度得到了有效提高。

6 结束语

本研究利用 SEED-DEC5502DSP 开发平台，实现了一个非特定人动态词表孤立词语音识别系统，该系统无需用户根据词条训练，移植性好，中央处理器采用定点 DSP 芯片，比采用浮点的芯片节省了成本。系统在软件设计中采用改进的实时端点检测算法，只对活动语音进行处理，从而减少了计算负担。在计算开销最大的维特比解码部分，笔者采用优化计算状态输出概率的方法在精度稍微下降的情况下，系统的解码速度得到了很大提高。实验结果表明，它对特定人孤立词的识别正确率达到 98%，具有一定的应用价值。

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. In: Proc ICASSP '00, 2000

2 Kao Y H, Rajasekaran P K. A low cost dynamic vocabulary speech recognizer on a gpp-dsp system. In: Proc ICASSP '00, 2000

3 杜利民等. 基于子词的嵌入式语音识别系统的研究. 电子与信息学报, 2005, 27(1)

大客户专线接入组网探讨

史训礼
（长沙市电信分公司　长沙 410007）

摘 要　如何优化网络结构、规范大客户专线的组网方式，是一个值得研究的课题。本文从大客户专线接入需关注的问题入手进行分析，通过对多种接入方式的比较，介绍了 MSAP（多业务接入平台）技术及其在大客户专线接入组网应用中的特点，提出并阐述了 MSAP 与 MSTP（多业务传输平台）有机衔接的大客户接入传输网络的组网思路。

关键词　专线接入；组网；多业务接入平台；多业务传输平台

1　前言

随着大客户专线接入业务越来越多，用户对专线接入业务的要求也不断提高，对电信运营商的接入网提出了更高的要求。目前大客户专线接入组网主要使用 PDH（准数字同步系列）光端机实现光纤接入，由于设备层级复杂、品牌多样，难以建立网

4 朱璇. 基于子词的嵌入式语音识别系统的研究. 清华大学博士学位论文，2003
5 王志强. 孤立词语音识别系统关键问题的研究. 北京邮电大学硕士学位论文，2004
6 ETSI Standard，ES 202 212v 1.1.1. Distributed speech recognition, speech processing, transmission and quality aspect, 2003
7 朱璇，李虎生，刘加等. 高性能汉语数码串快速识别算法的研究. 计算机研究与发展，2004，38（7）
8 Rogina F J. The bucket box intersection (BBI) algorithm for fast approximative evaluation of diagonal mixture Gaussians. In: Proc ICASSP.1996.

［作者简介］　张晨燕，石家庄经济学院信息工程学院副教授，主要研究方向：信号与信息处理、现代通信技术；孙成立，石家庄经济学院信息工程学院讲师，北京邮电大学在读博士，主要研究方向：语音信号处理与语音识别。

On Chip Realization of HMM Speaker-independent Isolated Word Speech Recognizer

Zhang Chenyan[1]，Sun Chengli[1, 2]
（1. School of Information Engineering, Shijiazhuang University of Economics, Shijiazhuang 050031,China;
2.Department of Information and Engineering, Beijing University of Post and Telecommunications, Beijing 100876,China）

Abstract　An embedded speaker-independent isolated word speech recognition system is designed and realized in the SEED-DEC5502 EVM platform. Compared with the speaker-dependent system, the speaker-independent recognition technique cannot requires training by the users and easy to use. With the help of a modified real time voice activity detection algorithm (VAD) based on the log-energy acceleration associated with voice onset, we only perform feature extraction and decoding to the active voice and ignore the frames of non-activity.To further decrease the computational loads, we analyze and optimize to the calculation of state output probabilities . Test on 100 words vocabulary shows that system provides a recognition accuracy rate of 98. 1% using only 1.03 times of real time .

Key words　speech recognition，embedded system，speech endpoint detect，state emission probability　（收稿日期：2006- 09- 13）