用 DSP 实现基于 VQ 的说话人识别系统

宋大杰，黄乡生，朱鹤树
（东华理工大学 机械与电子工程学院，江西 抚州 344000）

摘 要：在 TI 的 DSK5402 平台上构建了一个主要采用 VQ 方法的 6 个说话人识别系统。该系统采用了 10 阶的线性预测参数、10 阶的线性预测倒谱参数及基音参数，提出了一种改进的 LBG 算法，以避免在迭代过程中产生空胞腔，使之能适应多种距离测量。实验证明，本系统在指定文本的说话人闭集测试中取得了满意的效果。

关键词：数字信号处理器；说话人识别；矢量量化；LBG 算法

中图分类号：TN912.34 **文献标识码**：A **文章编号**：1674-7720(2012)05-0020-03

Realization of speaker recognition system based on VQ with DSP

Song Dajie, Huang Xiangsheng, Zhu Heshu
(Institution of Mechanical and Electronic Engineering, East China Institute of Technology, Fuzhou 344000, China)

Abstract: This paper implements a 6 speaker identification system on DSK5402 board of TI with VQ method. The system uses LPC, LPCC and pitch parameters. In order to avoid generating null partition in the iterative process, it proposes a modified LBG algorithm so that it can adapt to a variety of distance measurement. Experiments show that the system achieves satisfactory results on speaker identification test with appointed speech content.

Key words: digital signal processor; speaker recognition; vector quantization; LBG algorithm

自动说话人识别是一种自动识别说话人的过程，它着眼于提取语音信号中的个人特征，从而达到识别说话人的目的。说话人识别按是否规定说话人所说的内容可以分为文本有关型和文本无关型，前者要求待识别人说指定内容的一段话来进行识别，而后者对识别人说的内容无任何限制[1]。就整个说话人识别的发展来说，近几年说话人身份识别在理论和实验室条件下已经达到比较高的识别精度，并开始走向实际应用阶段。AT&T、欧洲电信联盟、ITT、Keyware、T-NETIX、Motorola 和 Visa 等公司相继开展了相关实用化研究，国内有关这方面的研究主要在中科院声学所、中科院自动化所、清华大学等研究所和大学中进行。本文采用 VQ 方法在 TI 的 DSK5402 平台上构建了一个文本有关的说话人身份识别系统，并采用线性预测语音合成方法来实现语音的人机交互。该系统具有使用方便、识别速度快和成本低等特点，具有广阔的应用前景。

1 算法的设计

本系统算法的流程如图 1 所示。首先将输入的经过数字化处理的语音信号进行预处理，然后提取其中与说话人有关的特征参数，接着对参数进行训练，为每个说话人生成一个模板。有了这组模板，在识别的时候，系统将提取新接收的语音的参数，并分别与这些模板进行比对，判断是否与某个模板匹配，最后给出判决结果。

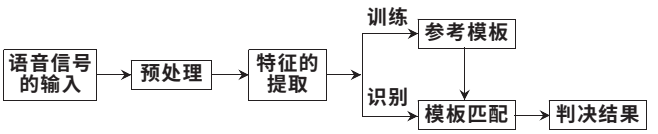

图 1 系统算法流程图

1.1 语音的预处理

本系统首先对采集到的语音信号进行预处理。这里认为待处理的语音是纯净的采样数字语音。预处理主要包括预加重、分帧和加窗、端点检测等操作。系统中采用一个 6 dB/倍频的一阶滤波器来进行预加重。为进行分帧和加窗，系统取帧长 10 ms(80 个样点)，窗长 30 ms(240 个样点、覆盖帧前 120 个样点、帧后 40 个样点)，由半个汉明窗和 1/4 个余弦窗组合而成。为减小计算量和提高计算精度，窗函数采用制表法，用浮点数算出数

值，再定点化为一张表以供调用。由于幅度门限法相对简单，计算量较小，因此系统采用它来进行语音的端点检测。通过预处理后，便可以应用短时分析技术逐帧提取出相应的特征参数。

1.2 语音特征参数的提取

本系统采用基音参数、LPC 参数和 LPCC 参数作为语音的特征参数。基音周期的估计采用自相关法，其具体过程是先求出一帧语音的自相关参数，然后系统在[20，39]、[40，79]、[80，143]3 个区间内各选一个自相关峰值点作为候选基音，接着对规格化的 3 个自相关峰值进行比较，选择最大的那个作为最终的基音。由于信号受声带共振峰特性的影响，求出的基音值会有所偏离，解决的办法是采用中心削波法，即将信号小于门限的点赋值为 0，大于门限的保持不变，然后将处理过的信号按以上方法求自相关。LPC 参数就是在线性预测（LP）分析中求得的全极点滤波器的系数集$\{a_i\}_{i=1}^{p}$，即在预测误差最小均方误差准则下，由公式

$$s(n)=-a_1s(n-1)-a_2s(n-2)-\cdots-a_ps(n-p) \quad (1)$$

可以得到 Yule-Walker 方程，然后采用 Levinson-Durbin 递推算法来高效地求解，该算法的时间和空间复杂度都比较低，适合本系统的定位。此外，系统应用了一个简单的利用 LPC 参数推导倒谱的方法，其递推公式为：

$$\begin{cases}\hat{s}(i)=-a_i-\sum_{j=1}^{i-1}\left(1-\frac{j}{i}\right)a_j\hat{s}(i-j) & 1\leqslant i\leqslant p\\ \hat{s}(i)=-\sum_{j=1}^{p}\left(1-\frac{j}{i}\right)a_j\hat{s}(i-j) & i>p\end{cases} \quad (2)$$

这种近似倒谱估计方法极大地节省了计算倒谱的运算量，并在语音识别的应用中取得了很好的效果，比较适合系统的定位要求，本系统在应用时取 $N_c=p$。

1.3 VQ 模式匹配

语音中的特征参数提取出来之后，就要对它们进行模式匹配（训练）以得到可供比对的模板。在模式匹配方面，由于 VQ 方法比较简单，实时性较好，实验证明，其在数据量较小的情况下比其他方法具有更好的鲁棒性，因此系统采用 VQ 方法[2]。VQ 方法中应用了 LBG 算法，本文使用了一种改进的 LBG 算法来避免传统 LBG 算法在迭代过程中产生空胞腔以及无法适用于多种距离测度等问题。其改进后的描述为：将样本点按现有码本进行空间划分得到一个胞腔 S 后，遍历 S 中的每一个元素 x_i，累加 S 内每一个元素与 x_i 的距离，相应距离记为 d_{sum_i}，该遍历的过程表示为：

$$d_{sum_i}=\sum_{j\in s}d(x_j,x_i) \quad i\in s \quad (3)$$

然后取使 d_{sum_i} 最小的那个 x_i 作为质心，即：

$$q=\arg\min_{i\in s}\{d_{sum_i}\} \quad (4)$$

最后按照传统的 LBG 算法进行迭代。

1.4 模式识别与判分机制

该系统的识别原理为：设系统要识别 N 个人的声音，分别提取各说话人训练语音中的特征参数 x，分别对各说话人的特征样本集$\{x_j\}_{j=1}^{L_1}$进行训练，得到 N 套用于识别的比对模板$\{x_j\}_{j=1}^{L_2}$。对于一段待识别的语音，首先提取出这段语音中相应的各帧特征参数，形成样本集$\{x_j\}_{j=1}^{L_2}$，然后分别用前面得到的 N 个码本对其进行矢量量化，取这 N 个平均量化误差最小的所对应的码本为可能匹配的码本。如果相应的平均量化误差小于事先设定的某个阈值，则认为此待识别的语音是那个码本对应说话人的语音，否则，则认为未找到匹配的说话人。

由于系统使用了 3 个参数进行判决，因此采用了一个判分机制来划分识别时的接受程度。具体方法是：首先按上述方法找到怀疑对象，分别记下 LPCC 参数和 LPC 参数的量化误差 $d_{rsltLpcc}$、$d_{rsltLpc}$ 以及对应的码本编号 rsltLpcc、rsltLpc，还要记下该语音段的基音到每个码本基音的距离。若 rsltLpcc≠rsltLpc，则认为该说话人不在闭集码本之内，退出判决；若 rsltLpcc=rsltLpc，则认为该段语音有可能匹配码本，记下 rsltLpcc 或是 rsltLpc 为 result，给 d_{result}^{lpcc}、d_{result}^{lpc}、d_{result}^{pitch} 各设定一个门限，分别记作 $GATE_{lpcc}$、$GATE_{lpc}$ 和 $GATE_{pitch}$。设置一个判决分量 $scale$，初始值设为 0，然后进行以下判别：如果 $d_{result}^{lpcc}<GATE_{lpcc}$，则 $scale=scale+2$；如果 $d_{result}^{lpc}<GATE_{lpc}$，则 $scale=scale+1$；如果 $d_{result}^{pitch}<GATE_{pitch}$，则 $scale=0$。这样 $scale$ 可能在 0、1、2、3 之间取值，它们分别代表：不匹配、不太可能匹配、有可能匹配、匹配。

1.5 线性预测语音合成

本系统中需要进行人机交流，如语音提示输入识别语句、识别结果提示等。为了节省系统资源，保存完整的语音提示信息是不现实的。由于线性预测语音合成具有占用资源少、数据率低和实现简单等特点[3]，而且系统交互所需要的语音对音质没有特别的要求，因此考虑用它来实现语音的人机界面。具体的实现过程是：首先利用 PARCOR 分析在 PC 上提取输入语音提示的 PARCOR 参数 k_m，以普通文本形式保存，然后在 DSP 平台上利用 k_m 由 PARCOR 分析的逆过程来实现语音提示的语音合成。语音分析的过程可以在 PC 上实现，不占用系统资源，而语音分析得到的参数保存为文本后大小仅有几 KB，与原始语音信号几百 KB 相比，占用系统数据区的资源少了很多，而语音合成的程序本身占用资源非常少，因此利用固定的参数文本和语音合成的办法实现有限的语音提示很适合本系统。

2 算法的 DSP 实现

在 TI 众多 DSP 产品中，TMS320C54X 系列用于多媒

体信号的处理及便携式设备，其片上资源及工作频率能满足一般的音频信号处理，而同样适合于多媒体处理的C64X和C62X系列虽然性能更加出色，但成本过高，因此本系统中采用TI的DSK5402集成开发环境作为开发平台。该平台上提供了一个PCM3002立体声编解码芯片，可以实现语音的采集和播放，通过它可以读入识别或是训练用的语音以及播放系统运行时所需要的语音提示命令。下面给出算法的具体优化方案。

2.1 精度保持与程序优化

本系统在信号处理过程中，由于迭代运算的大量出现，而TMS320C5402是定点DSP，为了防止误差的不断积累，需要在迭代的运算中做大量定点的高精度基本算术运算。为了在保持精度的同时又不过分降低运行速度，本系统将大量高精度的算术运算汇编化。本文根据自相关模块和LPC空间中求取IS距离模块自身的特点，着重对这两个模块进行优化。

自相关模块中输入的数据为加过窗的16 bit数组，输出数据为长字数组，其中归一化前采用32 bit，归一化后也采用32 bit，计算归一化数据所用除法采用Tayloer级数，使除法精度有效位达到32 bit(C++下浮点运算有效位为24 bit)。由于指数位对精度影响很小，因此这种方法下数据的精度已经超过了浮点运算。在汇编模块中使用了特殊指令EXP和NORM对数据进行位对齐，使保存未归一化数据时利用所有位，精度得到保持，使用累加、平方累加、块循环指令以加速程序运行[4]。通过优化，使得在保持精度的同时，对一帧信号作自相关耗费6 036个CLK，远远超过用C语言实现该模块的消耗。

LPC空间中求取IS距离模块的难点在于，向量本身是32 bit，中间计算都是48 bit，牵涉到32 bit×32 bit、48 bit×32 bit等高精度计算，而且该模块在训练和识别程序中都要反复被调用去计算LPC向量间的距离，对程序整体性能影响很大，只能将整块程序全部改为汇编，而在C语言中调用汇编的方法在速度上达不到要求。该模块中还使用了零开销循环、双字操作等指令来加速程序运行，而且利用汇编可以对存储器直接操作，使多个高精度共用一些存储器，避免了繁琐的赋值，节省了空间[4]。通过优化，计算一次IS距离只需4 211个CLK，而采用C语言中调用汇编需要13 054个CLK，由此可见，优化效果很明显。

2.2 实验结果与性能分析

实验采用长度为1 s的语音，VQ模式匹配的码本大小为16，对采样频率为8 kHz的单声道语音，采用10 ms的帧长逐帧提取参数，包括基音、10阶的LPC参数(相应的自相关参数)和10阶LPCC参数。按本文提出的VQ方法对单一说话人进行语音说话人识别。实验中首先训练这6个人的码本，所用的语音是“语音身份识别”，然后又使用这6个人另外一组相同语音进行鉴别，实验结果如表1所示。

表1 文本相关的说话人闭集实验结果

	Cbk0	Cbk1	Cbk2	Cbk3	Cbk4	Cbk5
Spk0	匹配					
Spk1		匹配				
Spk2			匹配			
Spk3				匹配		
Spk4					匹配	
Spk5						可能匹配

对每个人进行4~10次不等的重复试验并进行统计，结果表明，文本相关的说话人闭集实验的识别率在90%以上。本文分别对系统的运行时间及空间利用率进行了统计，如表2和表3所示。

表2 系统运行时间统计

操作	耗用CLK数	占用时间/s
一次参数提取	7 771 312	0.052
一次码本训练	68 275 383	0.428
一次码本识别	37 738 812	0.238

表3 系统的存储器使用统计

存储器使用	耗用字数	所占百分比/%
语音录入	0x2300	15
合成语音参数	0x17d6	10
VQ码本	0xde5	6
堆栈	0x3800	25
代码	0x1ed6	14
语音输出	0x2580	16
DSP\BIOS及其他	0x1f64	14
共使用	0xe455	100

由表2和表3可知，系统运行的速度较快，基本可以达到准实时的要求；整个系统占用近64 KB内存空间，其中约32 KB是必需的存储空间，32 KB是运行时所需的计算空间，由此可见，对系统资源的占用是较少的，完全可以满足系统的要求。

基于DSP的说话人身份识别系统具有精度高、适应性好、功耗低、费用低和体积小等优势，逐渐成为安全验证领域新的研究热点。本文在TI的DSK5402平台上构建了一个主要采用VQ方法的6个说话人识别系统，该系统在指定文本的说话人闭集测试中取得了满意的效果。与其他系统相比，本系统在实现算法上进行了改进，在保证识别率的同时提高了速度，具有更大的使用价值。

本文的主要创新点在于：在TI的DSP平台上实现了说话人身份识别算法的移植，并且在程序优化过程中针对系统算法中一些模块自身的特点，采取一系列手段使运算的精度得到保持、速度得到提高；系统还采用了线性预测语音合成方法来实现语音的人机交互界面，从而节省了更多系统内存，使用起来更加方便快捷。

参考文献

[1] 李财莲,赵小阳,王丽娟,等.说话人识别中关键技术的现状与发展[J].军事通信技术,2005,26(2):62.

[2] Huang H C, Pan J S, Lu Z M, et al. Vector quantization based on genetic simulated annealing [J]. Signal Processing, 2001, 81(7):1513-1523.

[3] 贺艳平.基于线形预测下的语音信号合成[J].西北民族大学学报(自然科学版),2010,31(80):43.

[4] Texas Instruments. TMS320C54X assembly language tools user's guide[Z]. 1997.

[5] 钱俊, 王芙蓉.C 代码在 TMS320C54X 上的手工汇编优化.DSP 专栏 [J]. 单片机与嵌入式系统应用,2004(5):71-72.

(收稿日期:2011-10-15)

作者简介:

宋大杰, 男,1985 年生, 硕士研究生, 主要研究方向:DSP 应用及开发、模式识别。

黄乡生,男,1950 年生,硕士,教授,主要研究方向:DSP 应用及开发、智能仪器与虚拟仪器的应用研究与开发。

朱鹤树,男,1984 年生,硕士研究生,主要研究方向:嵌入式应用及开发。