小词汇量语音识别与应用

凌 军[1,2] 兰 竞[2]

(重庆大学计算机学院 重庆 400044)[1](四川理工学院 自贡 643000)[2]

摘 要 本文分析了词汇量的语音识别原理和技术,对系统实现进行了软硬件功能划分,提出了硬件实现方案和软件实现方案,为应用产品的语音识别系统开发作了有益的探索。

关键词 语音识别,模式匹配,实现方案

1 引言

语音识别的研究工作可以追溯到20世纪50年代AT&T贝尔实验室的Audry系统,它是第一个可以识别十个英文数字的语音识别系统。但是直到60年代末70年代初期,语音识别才作为一个重要的课题展开工作,并且逐步取得实质性的进展。这是因为一方面计算机产业的迅速发展提出了使用要求,同时又提供了实现复杂算法的软、硬件环境。另一方面,更重要的是数字信号处理的理论和算法在那个时代取得了飞跃性的进展,如快速傅里叶变换、倒谱计算、线性预测算法、数字滤波器等。其中比较有代表性的是语音信号线性预测编码(LPC)技术的提出,以及日本学者将动态规划(DP)的概念用于解决孤立词识别时说话速度不均匀的难题,提出了著名的动态时间规整算法,简记为DTW(DYNAMIC TIME WARPING的缩写)。这有效地解决了语音信号的特征提取和不等长匹配问题。当词汇量较少以及各个词条不易于混淆时,DTW算法取得了很大成功,从而自60年代末期开始引起了语音识别的研究热潮。早期的语音识别系统大多是按照简单的模板匹配原理工作的特定人、小词汇量、孤立词识别系统。

识别技术经过多年研究,已经取得了重大的进展,克服了语音识别中非特定人、连续语音、大词汇量三大难题。

语音识别常用的算法有:基于模式匹配的动态时间规整法(DTW)、基于统计模型的隐马尔柯夫模型法(HMM)以及基于神经网络的识别法等。

在识别词汇量不是很大的应用场合中,如工业机器人、智能玩具等,基于模式匹配的语音识别技术不但简单方便、实时性好,而且有着较高的识别率,所以仍然有着广泛的应用前景。本文将讨论基于模板匹配的语音识别技术的实现。

2 语音识别原理

基于模式匹配的语音识别系统的原理如图1所示,语音识别一般分两个步骤。第一步是系统"学习"或"训练"阶段。这一阶段的任务是建立识别基本单元的声学模型以及进行文法分析的语言模型,即构建参考模式库。第二是"识别"或"测试"阶段。根据识别系统的类型选择能够满足要求的一种识别方法,采用语音分析方法分析出这种识别方法所需求的语音特征参数,按照一定的准则和测度与参考模式库中的模型进行比较,通过判决得出结果。

2.1 预处理

语音信号分析是语音信号处理的前提和基础,只有分析出代表语音信号本质特征的参数,才有可能利用这些参数进行高效的语音识别处理。根据分析方法不同,可将语音信号分析分为模型分析法和非模型分析法。对于小词汇语音识别系统采用依据模型进行分析的线性预测分析是比较好的。在对语音信号分析之前应将语音信号进行前端处理,其中包括语音的滤波、数字化等预处理,以便为语音信号特征提取和语音识别打下基础。

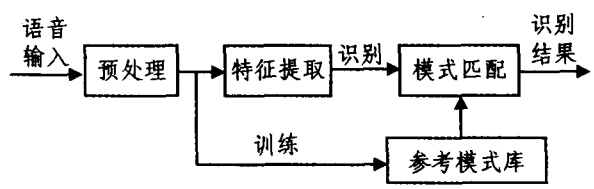


图1 基于模式识别的语音识别原理图

滤波的目的有两个:(1)抑制输入信号各分量中频率超出f/2的所有分量(f为采样频率),以防止干扰。(2)抑制50Hz的电源工频干扰。滤波器必须是一个带通滤波器。为了减少硬件设备,可以采用数字滤波器。

A/D转换器是将原始的模拟语音信号变为数

字信号，从而得到时间和幅度上均为离散的数字语音信号。A/D转换时采样频率的选择很重要，它关系到采样过程中是否会丢失信息，在语音信号处理中，采样频率通常为7～10kHz。由此可见，选择的A/D转换器性能的好坏对语音信号的处理也是很重要的。

预处理一般包括预加重、加窗、分帧和端点检测等。

预加重的目的是提升高频部分，使信号的频谱变得平坦。保持在低频到高频的整个频带中，能用同样的信噪比求频谱，以便进行频谱分析或声道参数分析。加窗分帧的目的是为了利用有限容I的数据区依次处理数f极大的语音数据。根据语音信号在10～20ms内语音信号特性不变的特点，一般取顿长为20ms，帧移为帧长的0～1/2倍。经过加窗分帧处理后，语音信号就已经被分割成一帧一帧的加窗函数的短时信号，然后再把每一个短时语音帧看成平坦的随机信号，利用数字信号处理技术来提取语音特征参数。在进行处理时，按帧从数据区中取出数据，处理完成后再取下一帧。最后得到由每一帧参数组成的语音特征参数的时间序列。

端点检测是将语音信号从背景噪声中提取出来，以确定语音信号的起止点。常用的方法有短时过零率和短时平均能量等几种。

2.2 特征提取

经过预处理后的语音信号，要对其进行特征提取，即特征参数分析。该过程就是从原始语音信号中抽取能够反映语音本质的特征参数，形成特征矢量序列。目前语音识别所用的特征参数主要有两种类型：线性预测倒谱系数(LPCC)和美尔频标倒谱系数(MFCC)。LPCC系数主要模拟人的发声模型，未考虑人耳的听觉特性。它对元音有较好的描述能力，而对辅音描述能力差。其优点为计算量小，比较彻底地去掉了语音产生过程中的激励信息，易于实现。MFCC系数考虑到了人听觉特性，并具有很高的鲁棒性和抗噪声能力，但因为提取MFCC参数要在频域处理，计算傅立叶变换将耗费大量宝贵的计算资源。语音特征提取是分帧提取的，每帧特征参数一般构成一个矢量，因此，语音特征是一个矢量序列。该序列的数据率一般可能太高，不便于其后的进一步处理，为此，有必要采用很有效的数据压缩技术方法对数据进行压缩。矢量量化就是一种很好的数据压缩技术。

2.3 参考模式库

参考模式库是将一个或多个说话者的多次重复的语音参数经过训练得到的。它是声学参数模板。建立参考模式库是在系统使用前获得并存贮起来的。参考模式库的建立的过程称为训练过程。

2.4 模式匹配

模式匹配是将输入的待识别的语音特征参数同训练得到的参考语音模式进行逐一比较分析，获得最佳匹配的参考模式便为识别结果。目前常用的语音识别算法主要有：动态时间规则、离散隐马尔可夫模型、连续隐马尔可夫模型、人工神经网。

3 小词汇量语音识别系统的软硬件方案

3.1 硬件框图

语音识别系统硬件框图如图2所示。

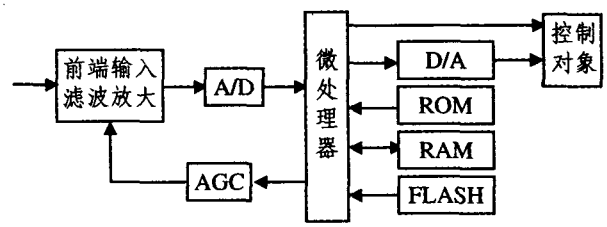


图2 语音识别系统硬件框图

小词汇量语音识别系统对存储器容量要求不高，一般的小规模集成芯片就可满足。在硬件系统中，前端滤波是为了消除干扰和可能造成混淆的成分。由于语音信号是较弱的信号，因此，要对输入的语音信号进行放大，为了使语音信号的放大值控制在一定的范围内，在硬件中设有自动增益控制AGC。根据受控对象的性质，可直接将数字信号传给它，或经过D/A将数据量信号转换成受控对象所需的模拟量信号传给它。RAM用于存储提取的语音特征参数，而语音提示部分可固化在ROM中或存储在FLASH内。语音识别程序也被存储在ROM或FLASH中。控制对象是根据此系统所应用的领域不同而有所区别，可以接收模拟信号，也可接收数字信号。在设计时，依据对控制对象的不同，还要考虑相应驱动电路。

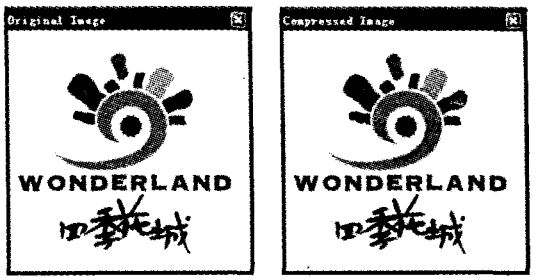
对于应用系统而言，其硬件组成有许多其它因素需要考虑。首先是成本，由于成本的限制，一般使用定点DSP，有时甚至只能考虑使用MPU，这意味着算法的复杂度受到限制；其次，应用系统对体积可能有一些限制，如果受限，则需要一个高度集成的硬件平台。

一种理想的硬件组成是系统级的集成芯片。它不只是把功能复杂的若干个数字逻辑电路放入同一个芯片，做成一个完整的单片数字系统，而且在芯片中还应包括其它类型的电子功能器件，如模拟器件和存储器。

3.2 软件实现

语音识别系统通过硬件开关实现训练和识别的转换。其工作过程如图3所示。语音识别系统开机

(下转第159页)



(a) (b)

图 4 小波压缩的原图以及压缩后的效果图
((a)、原图 (b)小波压缩后的图)

以上是将信噪比设置成 35.202dB 后，图像由原先的 383448bit 下降到 50914bit 的效果图，压缩比例达到了 86.7%，有比较好的压缩效果。

结论 JPEG2000 压缩算法中可以选择的压缩比的功能可以很好地解决图像大小和图像质量之间的协调问题。本文提到的仅仅是 JPEG2000 算法中的一部分，JPEG2000 中还有一个更加重要的功能本文没有涉及到，即选择“感兴趣区域 ROI”，该功能允许对比特流的任意访问和处理，用户可以把图像中的特定部位定义为重要区域，并对这个区域进行无损压缩。对于图像中的其它区域，则可以适当调整压缩比以减小图像的数据量，这样既可以保证特定区域的图像质量，又可以提高压缩比。

综上所述使用 JPEG2000 压缩图像有以下优点：(1)在较高的压缩比下可以保持好的图像质量。(2)具备有损压缩和无损压缩的可选择性，可以同时满足各种需求。(3)感兴趣区域的选择和渐进传输的功能更是其他压缩方法所不具备的。因而 JPEG2000 压缩方法非常值得推广应用，但下一步需要对小波压缩的速度和效果需要进一步的分析，根据不同环境进行选择和优化。

参 考 文 献

1 刘芳敏，等. JPEG2000 图像压缩过程及原理概述. 计算机辅助设计与图形学学报，2002，10-1：905～916
2 刘芳敏，等. 基于小波变换的 J PEG2000 算法及其在医学图像压缩中的应用. 医疗设备信息，2005. 12～13
3 林济南，等. 在医学图像 DICOM 格式中实现 JPEG 压缩算法. 北京生物医学工程，2004，23 (3)：209～211
4 Taubman D S, Marcellin M W. JPEG2000：Image Compression Fundamentals，Practice and Standards M . Massachusetts：Kluwer Academic Publishers，2002. 255～258
5 Bil gin A. Compression of Electrocardiogram Signals using JPEG2000. Transactions on Consumer Electronics，2003，9 ：833～840
6 刘俊主编 . Delphi 数字图像处理及高级应用 . 北京：科学出版社，2003. 9

(上接第 147 页)

就检测环境声音，若超过一个设定的阈值，则认为检测到端点，开始采集信号，否则继续检测环境声音。

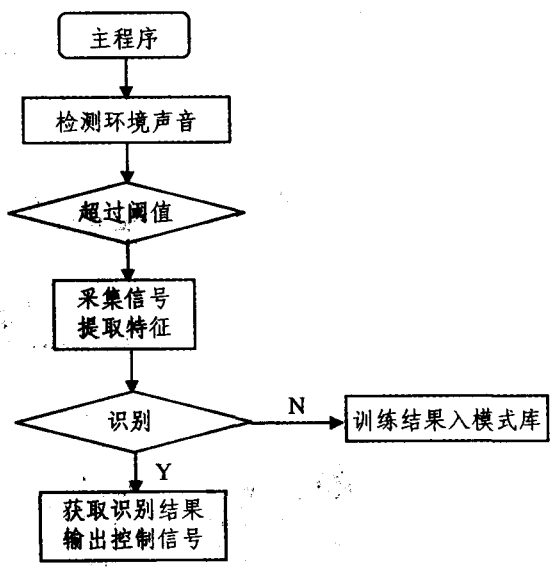


图 3 软件工作流程框图

若是训练阶段，则采集信号并提取特征后按预先制定的方式存入模式库；若是正常工作阶段，则将提取的特征与模式库中的样板进行比较，获取识别结果，用以控制外部的对象。

结论 基于模式匹配的小词汇量语音识别，虽在理论上并不是很前沿，但其简单、易实现的优点，在具体的应用领域，如实用机器人、智能玩具等的开发和应用，具有很高的实用价值。系统的软硬件功能分配已很明确，但不同的应用场合对系统的体积、响应速度有不同的要求，可在开发产品的时候作更为详细的分析，以确定具体的硬件实现方案。

参 考 文 献

1 胡光锐. 语音处理与识别[M]. 上海：上海科学技术文献出版社，1994
2 〔美〕拉宾纳 LR，谢弗 RW 著 . 语音信号数字处理[M]. 朱雪龙译. 北京：科学出版社，1983
3 易克初. 语音信号处理[M]. 北京：国防工业出版社，2000
4 赵文，杨没宇，杨鉴. 孤立词识别口. 计算机应用，2001，21(6)：12～14
5 井贞熙. 朱家新译. 数字语音处理. 北京：人民邮电出版社，1985 (9)：2～95