基于 SoPC 的孤立词语音识别系统的设计

孙 玉，郭宝增
（河北大学 电子信息工程学院，河北 保定 071002）

摘 要：采用 SoPC 方法，实现了基于动态时间规整（DTW）算法的孤立词语音识别系统，该系统可以作为电器系统的语音命令控制模块使用。考虑嵌入式系统的特点，对端点检测算法和模式匹配算法进行了选择和调整。实验表明，该语音识别系统运行速度和识别准确性能够适应语音控制的要求。SoPC 设计方式灵活，适合对系统进行改进升级。

关键词：SoPC；Nios II；语音识别；动态时间规整
中图分类号：TP391 **文献标识码**：A **文章编号**：1674-7720(2012)02-0074-03

Design of isolated-word speech recognition system based on SoPC

Sun Yu, Guo Baozeng
(College of Electronic and Informational Engineering, Hebei University, Baoding 071002, China)

Abstract: This system uses the SoPC method to realize an isolated-word speech recognition system. The system is based on the improved DTW algorithm, and it can be easily integrated into other systems. The endpoint detection algorithm and pattern recognition algorithm are selected and modified. The speed and accuracy of the system can satisfy the needs of voice control. By SoPC technology, the system is flexible and easy to be modified for different applications.

Key words: SoPC; Nios II; speech recognition; DTW

随着计算机技术、模式识别技术等的发展，国内外对语音识别的研究也不断进步。目前电器、家居智能化的实际需求使得语音识别技术成为一个研究热点。例如，美国约翰·霍普金斯大学语言和语音处理中心多年来一直致力于推动语言和语音识别的研究和教育，CLSP 每年一度的夏季研讨会对语音识别的各个领域都产生了深远的影响。国内，中国科学院等也在语音识别领域有较大进展。

相对于基于 PC 机平台的大词汇量语音识别系统，嵌入式系统中要求语音控制模块占用资源少，功能简洁，可作为独立的语音识别系统或其他系统的语音控制部分。因此，根据语音识别系统的准确性、实时性的要求和 SoPC 实现方式的特点，在介绍实现该语音识别系统的基本流程的基础上着重探讨以下两部分内容：(1)由于端点检测算法对识别的准确性影响较大，本系统探索适合 SoPC 设计的端点检测算法，从而使得系统的识别准确性有所改进；(2)模式匹配时，对同一模板采用了多个局部判决函数，求多个累加总距离的平均值作为最终的判决依据，进一步提高了识别结果的可靠性。

可编程片上系统 SoPC(System on Programmable Chip)是 Altera 公司提出的一种基于 FPGA 的嵌入式系统解决方法，采用软硬件结合设计的思想，实现方式简单灵活[1]。设计中采用高性价比的 EP2C70 FPGA 芯片。实验结果表明，系统运行良好，能够满足中、小词汇量孤立词语音识别系统的要求。

1 设计方案

语音识别系统的逻辑流程如图 1 所示。采样得到的语音信号要经过预处理、端点检测、特征参数提取，然后根据用户指定的工作模式（识别模式或训练模式），进行模式匹配并输出识别结果，或者训练得到该词条的模板，并存入模板库。因此，在硬件资源允许的条件下，用户可以自定义训

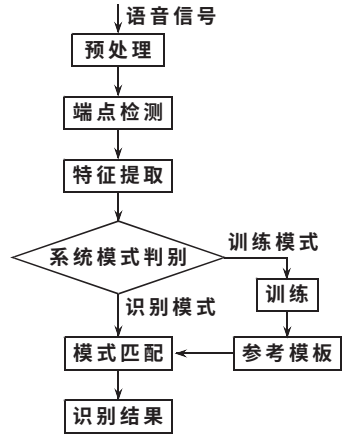

图 1 语音识别系统逻辑流程图

练模板，更新模板库，拓展系统的应用范围。

1.1 预加重和端点检测

系统采用 8 kHz 采样，由音频编/解码芯片 WM8731 采样得到的语音数据，经过 FIFO 数据缓存器传输到系统的 SDRAM 中，然后对 SDRAM 中的数据进行后续处理。设定 256 个采样点作为一帧，每个孤立词采集 100 帧(3.2 s)数据。

(1)预加重：处理的第一步要对采集到的数字语音信号进行预处理，主要是预加重。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相同的信噪比求频谱。通过一个滤波器对信号进行滤波，滤波器的传递函数为：

$$H(z)=1-0.98z^{-1} \tag{1}$$

(2)端点检测：从数字语音信号中快速有效地切分出语音段，对于整个系统的识别速度和识别准确性影响较大。根据汉语语音的特点，一般一个汉语单词的开始部分是清音，接下来是浊音，清音较弱，浊音较强。因此在端点检测部分，采用了基于短时能量和短时过零率的双重检测。首先根据浊音粗判起始帧，然后根据清音，细判起始帧。语音的起始帧和终止帧都是经过粗判和细判之后得出，从而保证端点检测的准确性[2]。

短时能量由下式计算：

$$E(i)=\sum_{n=0}^{N-1}x_i^2(n) \tag{2}$$

短时过零率由下式计算：

$$ZCR(i)=\sum_{n=0}^{N-1}|x_i(n+1)-x_i(n)| \tag{3}$$

其中，i 为帧标志（$i=0,1,2,\cdots,100$）；N 为每帧的采样点数，此处设定 $N=256$；$x_i(n)$ 表示第 i 帧中第 n 个采样点的值。因此，$E(i)$ 为第 i 帧的短时能量，$ZCR(i)$ 为每帧的短时过零率，对每帧求一次短时能量和过零率。

端点检测部分程序流程如图 2 所示。

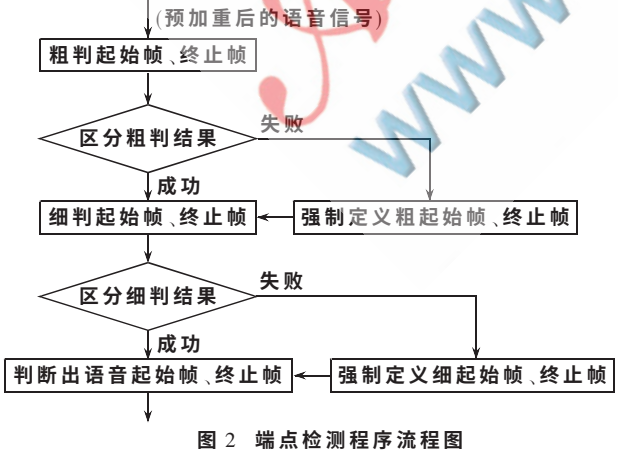

图 2 端点检测程序流程图

1.2 特征提取

经过预加重和端点检测之后得到语音段采样值构成的向量序列。接下来对该向量序列进行特征参数分析，目的是提取合适的语音特征参数，使特征向量序列在语音识别时，类内距离尽量小，类间距离尽量大。特征参数的提取同样是语音识别的关键问题，特征参数的选择直接影响到语音识别的精度。结合 SoPC 设计的需求，选择提取语音信号的美尔特征参数(MFCC)[3]。MFCC 能够较好地反映人耳的听觉特性。

为求识别系统简洁，每词条固定采集 3.2 s 的语音信号，采样频率为 8 kHz，经端点检测切分出语音段，然后将语音段进行分帧(每帧 256 个采样点)，每帧提取一组 14 维的 MFCC 参数，组成一组特征参数向量序列，作为待识别语音段的特征参数。

1.3 模式匹配

对于大词汇量的非特定人语音识别系统，模式匹配多采用基于模型参数的隐马尔可夫模型(HMM)的方法或基于非模型参数的矢量量化(VQ)的方法。但是 HMM 算法模型数据过大，对存储空间和处理速度的要求高，不适合嵌入式系统。VQ 算法虽然训练和识别的时间较短，对内存要求也较小，但识别性能较差。因此考虑到嵌入式系统系统资源有限以及运算能力限制，而又需要保证识别准确性，决定采用基于动态时间规整的算法(DTW)进行模式匹配。

由于每个人的发音习惯不同，以及同一个人每次说同一个单词时说话速度具有随机性，因此会导致每次采样得到的语音数据序列长度具有随机性。DTW 算法由日本学者板仓(Itakura)提出[4]，能够较好地解决语音识别时单词长度具有随机性这一问题。

DTW[5-6]算法将时间规整和距离测度计算相结合，描述如下：

(1)将特征提取部分提取出来的特征向量序列与模板库中每个词条的特征向量序列逐帧计算距离，得到距离矩阵。对应帧之间的距离是两帧的特征向量中对应分量的差值的平方和。距离矩阵中元素的计算式为：

$$D[i][j]=\sum_{m=1}^{K}(x[i][m]-y[j][m])^2 \tag{4}$$

其中，$D[i][j]$ 为距离矩阵的元素，表示待识别语音段特征向量序列第 i 帧和该条参考模板向量序列第 j 帧之间的距离，$i=0,1,2,\cdots,I-1$；$j=0,1,2,\cdots,J-1$。I、J 分别为待识别语音的特征向量序列和该条参考模板序列的总帧数。$x[i][m]$ 为待识别语音的特征向量序列第 i 帧向量的第 m 维分量，$y[j][m]$ 为该条参考模板的第 j 帧向量的第 m 维分量。K 为对切分出的语音段每帧语音的原始采样数据提取的特征向量的维数，该识别系统中每帧提取 14 维的 MFCC 参数，因此 $K=14$。

(2)按照一定的局部判决函数，由距离矩阵计算出累加距离矩阵(求得的累加距离矩阵最末一个元素的值即为待识别语音和该条参考模板之间的总距离)，得到累加距离矩阵的同时得出最佳规整路径[3-4]。图 3 所示为设计中采用的三种局部判决函数。

(a)基本对称型 (b)拓展类型 (c)拓展类型

图3 局部判决函数

在如图3(a)所示的基本对称局部判决函数下，累加距离矩阵中元素的计算公式为：

$$G[i][j]=D[i][j]+\min[G(i-1,j),G(i-1,j-1),G(i,j-1)] \quad (5)$$

其中，$G[i][j]$为累加距离矩阵的元素，$i=0,1,2,\cdots,I-1$；$j=0,1,2,\cdots,J-1$。I、J分别为待识别语音特征向量序列和该条参考模板序列的总帧数。$G[i][j]$的值等于在选定的局部判决函数下，从距离矩阵中前一位置到达距离矩阵中的第i行、第j列交汇处($D[i][j]$处)的最小代价值与距离矩阵中该位置的元素值$D[i][j]$之和，这里规定$G[0][0]=D[0][0]$。

在基本对称局部判决函数下，有三条路径(垂直路径、对角线路径和水平路径)到达距离矩阵中的第i行、第j列交汇处($D[i][j]$处)，其代价值分别为$G[i-1][j]$、$G[i-1][j-1]$和$G[i][j-1]$。取三者中的最小值作为该步转移的最小代价值。最小的代价值对应的路径为该步的最佳规整路径。求得的累加距离矩阵的最末一个元素$G[I-1][J-1]$的值就是待识别语音和该条参考模板之间的累加总距离。

通常计算累加距离矩阵的局部判决函数采用一种固定的形状，这样对同一个累加距离矩阵只能计算出一个总距离，未能较全面地统计出待识别序列与参考模板之间的距离关系。因此，尝试对同一个距离矩阵采用多个不同的局部判决函数(如图3所示)计算出相应个数的累加距离矩阵。最后，以多个累加总距离的统计平均值作为待识别语音与该条模板的距离。

程序中采用上述三种局部判决函数。在各种局部判决函数下求解累加距离矩阵各元素的计算公式分别为：

$$G_{(a)}[i][j]=D[i][j]+\min(G(i-1,j),G(i-1,j-1),G(i,j-1)) \quad (6)$$

$$G_{(b)}[i][j]=D[i][j]+\min(G(i-2,j-1),G(i-1,j-1),G(i,j-1)) \quad (7)$$

$$G_{(c)}[i][j]=D[i][j]+\min(G(i-2,j-1)+G(i-1,j)),G(i-1,j-1),(G(i-1,j-2)+G(i,j-1)) \quad (8)$$

其中规定$G_{(a)}[0][0]=G_{(b)}[0][0]=G_{(c)}[0][0]=D[0][0]$。

上述三个累加距离矩阵中最后一次求得的元素的值，即为采样相应的局部判决函数时，待识别语音序列与该模板序列之间的总距离，分别为：$G_{(a)}[I-1][J-1]$、$G_{(b)}[I-1][J-1]$和$G_{(c)}[I-1][J-1]$。取三个总距离的统计平均值，可得待识别语音与该条模板之间的统计平均距离：

$$G[I-1][J-1]=(G_{(a)}[I-1][J-1]+G_{(b)}[I-1][J-1]+G_{(c)}[I-1][J-1])/3 \quad (9)$$

同理，对待测语音与模板库中的每条模板求得一个统计平均距离。由判决逻辑判断出各统计平均距离值中的最小值，相应的模板所指向的单词即为最终的识别结果。

2 系统实现

系统硬件部分如图4所示，包括FPGA芯片、Flash、SDRAM、音频编解码芯片WM8731、按键以及LCD1602。在FPGA芯片中添加NiosII软核CPU，并建立片外Flash、SDRAM、音频编解码芯片WM8731和LCD1602的接口部分。

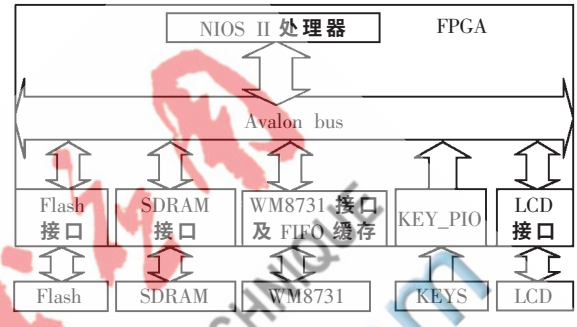

图4 系统硬件结构

系统中的Flash芯片用于存储FPGA硬件设计文件、系统初始化所需代码和用户程序以及参考模板库，SDRAM作为系统内存。系统上电后，硬件编程文件以及用户程序读入SDRAM中，并且初始化各硬件设备。初始化完毕后，如果选择系统进入识别模式，CPU通过I²C总线向WM8731写入命令，控制音频芯片WM8731采集语音数据并进行A/D转化，得到的数据经过WM8731的数据输出端口在FPGA内部构建的FIFO缓存，存入内存中，然后对数据进行后续处理并进行识别，最终输出识别结果。

系统采用SoPC方法实现，按需要定制硬件模块，设计过程简洁灵活，对系统的后续升级维护也比较方便。且实验表明：(1)采用双重判决机制进行端点检测，能够得到较好的识别效果；(2)采用较简单的DTW方法进行模式匹配，一方面在消耗系统资源较少的情况下能够得到较高的识别速度，使得整个系统能够满足实时性的要求。另一方面，在模式匹配时采用不同的局部判决函数求得多个累加总距离，进而求得累加总距离的统计平均值。以统计平均值作为最终的判决依据，相对于只采用单一局部判决函数下得到的识别结果理论上可靠性更高。

参考文献

[1] 周立功.SoPC嵌入式系统基础教程[M].北京：北京航空航天大学出版社，2008.

[2] 刘华平，李昕，徐柏龄，等.语音信号端点检测方法综述及展望[J].计算机应用研究，2008(8)：2278-2283.

[3] 罗希，刘锦高.基于NIOS的ANN语音识别系统[J].计算机系统应用，2009(12)：144-146.

[4] 赵力.语音信号处理[M].北京：机械工业出版社，2003.

[5] 王娜，刘政连.基于DTW的孤立词语音识别系统的研究

与实现[J].九江学院学报(自然科学版),2010(3):31-39.
[6] 姚天任.数字语音处理[M].武汉:华中科技大学出版社,1991.

(收稿日期:2011-09-29)

作者简介:

孙玉,男,1987年生,研究生,主要研究方向:语音识别和FPGA。

郭宝增,男,1953年生,教授,主要研究方向:集成电路设计。