语音识别在 MP3 播放器上的应用

李兰芳[1],汪道辉[1],董海疆[1],李国清[2]
(1.四川大学 自动化系,四川 成都 610065;
2.桂林电子科技大学 电子工程系,广西 桂林 541004)

摘 要:介绍了基于 DSP 的语音识别在 MP3 播放器上的应用。对语音识别部分软硬件进行了设计,通过对语音信号的识别,实现语音控制 MP3 播放器的操作。

关键词:语音识别 控制 MP3 DSP

随着社会的发展以及技术的革新,DSP 的应用越来越广泛,这与 DSP 越来越高的性价比不无关系。DSP 被广泛应用于数字信号处理、语音识别及处理、图形/图像处理、自动控制和通信、仪器仪表检测、医学、家电等多方面。本文将介绍一种基于 TMS320C55X 的语音识别在 MP3 播放器中的应用。

1 语音识别的应用

语音识别的应用较广泛,其主要应用在两个方面:(1)将语音输入识别后存储,不产生具体的控制功能;(2)将语音输入识别后形成具体的控制指令,达到具体的控制目标。前者可应用于录音笔等语音输入设备,录音后自动产生文稿存储,为广大新闻工作者的工作提供了方便。后者可应用于家电、医疗等方面,语音可以解决具体的需要手动控制的问题,这将大大地有益于人们的工作和生活。本文将结合语音识别第二个方面的应用,提出一种应用于 MP3 播放器的方案。

2 设计目标

语音识别在 MP3 播放器中的应用主要利用 DSP 在语音识别方面的优势,在产品上附加了语音控制的功能,可设计为对特定人的语音识别,在语音识别模式下没有特定人的语音,无法开启 MP3 播放器。这既减少了配件方面的要求也可以防盗。当然,为保证此功能,在系统学习语音时需要密码。

由于 MP3 播放器的很多功能较成熟,在此不作赘述,本文只介绍语音识别及处理与 MP3 播放器的结合。

3 语音识别系统

语音识别系统框图如图 1 所示。系统由三个部分构成:

(1)语音输入部分。通过 MIC 采用差分输入的方式,以消除部分干扰。

(2)语音信号的数字化和预处理。语音信号的数字化一般包括放大及增益控制、反混叠滤波、采样、A/D 变换及编码,通过端点检测,把语音部分区分出来。预处理

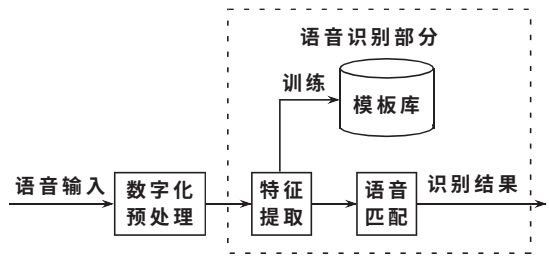

图 1 语音识别系统框图

一般包括预加重、加窗和分帧等。

(3)语音识别部分。根据在训练模式下建立的模板库对输入的语音信号进行识别。通过模板匹配,识别出具体的语音结果,对应于具体的控制信号,输出控制 MP3。

4 语音识别控制部分的硬件结构

语音识别控制部分的硬件框图如图 2 所示。它由语音信号处理模块和系统控制模块组成。

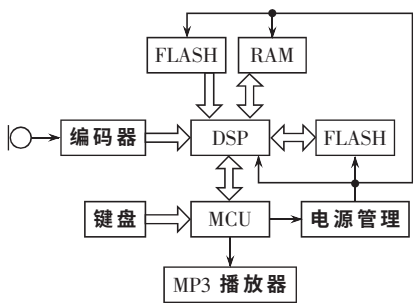

图 2 语音识别控制部分的硬件框图

语音信号处理模块由 DSP 芯片、两片 FLASH、编码器、RAM 组成。其中 DSP 是整个语音识别模块的核心,负责语音识别以及 FLASH 的读写控制。DSP 芯片的优点是运算速度快、内存空间大、数据交换速度快,可用来实现复杂的算法,提高识别率,减小反应延时,得到较高的识别性能。一片 FLASH 用于 DSP 芯片的引导程序的存放,定义为只读;另外一片 FLASH 用于样本集的存储,定义为可读写。编码器主要完成模/数转换和滤波功

能。语音识别需要很大的 RAM 空间，作为 DSP 芯片的 RAM 扩充。

系统控制模块负责整个遥控器的系统控制，它由单片机、键盘、控制信号发生器和电源管理电路组成。单片机作为主控芯片，主要对键盘进行扫描；控制 DSP 芯片进行语音训练、识别；将识别结果转换成相应的控制信号，输出控制 MP3；同时它也对电源进行管理。

5 TMS320C55X 系列简介

TMS320C55X 系列具有性能高和功耗低的特性，性能较 TMS320C54X 提高了 5 倍而功耗只有其 1/6，满足声音识别的需要，不会给便携型的 MP3 播放器带来任何不便。TMS320C55X 有模数转换器 ADC、通用串行总线 USB、多媒体卡控制器 MMC 等丰富的外设，方便扩展。TMS320C55X 的寻址空间达 16MB，能够运行较大的程序和足够多的语音样本。以 5510 为例，其时钟速率为 160/200MHz，160K×16 位片内 RAM、16K×16 位片内 ROM、三个通道缓冲串行口、32 位外部存储器接口。

6 语音识别控制部分软件设计

语音识别控制部分的功能主要为训练功能和识别功能。

训练功能在训练模式下建立模板库。识别功能将语音信号识别为具体的控制信号，将其传输并控制 MP3 播放器，并更新其模板库。

训练功能和识别功能的程序流程图如图 3 所示。

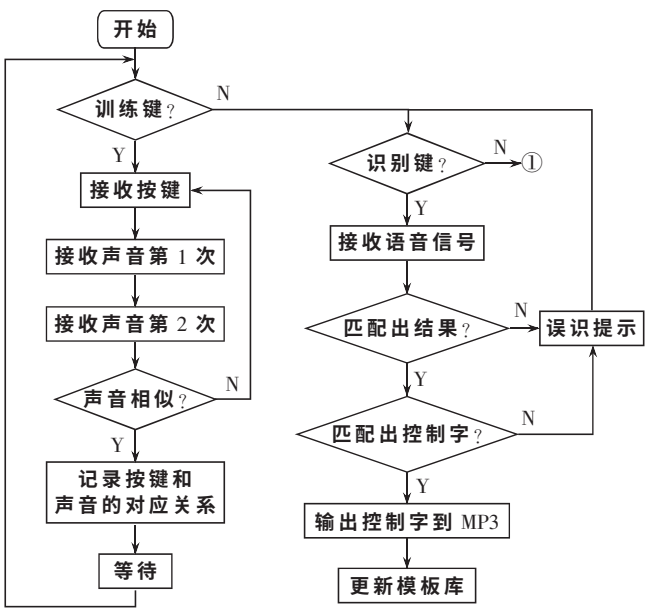

图 3　训练功能和识别功能程序流程图

图 3 中，①指向输入密码后手动操作。本文不对其手动部分作程序流程图分析。

7 语音识别算法

语音识别的软件实现主要从数字化和预处理、端点检测、特征提取以及模板匹配等方面介绍。

(1)数字化和预处理

预滤波器是一个带通滤波器，系统可选择 3.4kHz、60Hz 分别为上、下截止频率，采样频率为 8kHz，12 位的采样精度。这样的选择已经可以防混叠，当然在要求较高的情况下可以扩大其滤波器的带宽和使用更高的采样频率。

预加重可以提升高频部分，使信号频谱变得平坦，便于进行频谱分析或者声道参数分析。它一般可选一阶的数字滤波器：

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

式中，μ 的值小于且接近 1，这里为 0.98。

(2)加窗处理

语音信号可以看作短时平稳的语音信号，从一个长时间范围看是非线性的，但是在 10~30ms 的时间范围内可以认为是线性平稳。所以选择一个有限长度的窗函数对其进行截取，并使其在信号序列上滑动，从而可以处理整个信号序列。

常用的窗函数有矩形窗、汉明窗等。这里选择汉明窗，因为汉明窗的带宽和带外衰减都比矩形窗大，且它保存了高频部分(即波形)的细节。

$$h(n)=\begin{cases}0.54-0.46\cos[2\pi n/(N-1)] & 0\leqslant n\leqslant N-1\\ 0 & n\text{ 为其他}\end{cases}$$

系统可以选 N 为 100~200 个点，这里选择 $N=160$，滑动的幅度为 80 个点。加窗后，语音信号可以看成是由很多帧的短时平稳信号组成。

(3)端点检测

从背景噪声中区分出语音信号的起止点，除了可以大大减少计算量及运算时间外，还可以提高识别的准确性。这里采用短时平均幅度函数和平均过零率两种方法分析。

①短时平均幅度法。短时平均幅度函数是一帧语音信号能量大小的表征，其定义如下：

$$M_n=\sum_{m=0}^{N-1}|x_n(m)|$$

其中：$x_n(m)$为数字语音信号样本，$N=160$。根据噪声环境，取一门限值，M_n 高于此门限值则为浊音段，否则为清音段。

②短时过零率。它表示一帧语音信号改变符号的次数，定义如下：

$$Z_n=\frac{1}{2}\sum_{m=0}^{N-1}|\mathrm{sgn}[x_n(m)]-\mathrm{sgn}[(m-1)]|$$

其中 $\mathrm{sgn}[x]=\begin{cases}1,(x\geqslant 0)\\ 0,(x<0)\end{cases}$

由于浊音段的过零率比清音段低得多，可以此区分。

(4)语音信号的特征提取

采用 LPC 倒谱特征系数 LPCC 进行特征提取，它可以用于说话人和说话内容的识别。这里先求 LPC 的系

(接上页)

数，再根据两者的关系确定 LPCC 的参数。线性预测分析，即用过去 p 个信号的样点值预测现在或未来的样点值。线性预测器定义如下：

$$\hat{s}(n)=\sum_{i=1}^{p} a_i s(n-i)$$

其中，p 为 LPC 的阶数，这里可选 $p=12$；a_i 为相应的 LPC 系数。通过相应公式的转换，LPC 系数可以转换成 LPCC 系数，选择 LPCC 系数的阶数为 16。

(5)模板匹配

将输入语音的特征矢量序列依次与模板库中的每个模板进行相似度比较，其中相似度最高的作为识别结果的输出。这里用到动态时间归整技术(DTW)。它是把时间归整并与距离测度计算结合的一种非线性归整技术，它解决了同一个人在不同的场合说的同一句话的同一个音，时间长度不可能完全相同的问题。它用满足一定条件的时间归整函数描述待识别的模式和参考模板的时间对应关系，求解两模板匹配即累计距离最小时所对应的归整函数。因而，它保证了两模板间存在的最大的声学相似性。

通过语音识别的应用，可以使产品以及人们的生活增添很多内涵。应用 DSP 技术，将语音识别与 MP3 结合只是一个很小的应用。语音识别的应用前景很宽阔。

参考文献

1 赵 力.语音信号处理[M].北京：机械工业出版社，2003

2 朱纯益，陆建华，刘润生.基于 DSP 的声控电子记事本的设计与实现.电子技术应用，2002；28(9)：71~73

3 马洪连，朱 杰，杨凤岐等.基于 DSP 的声控系统的设计与实现.测控技术，2005；24(12)：30~32

(收稿日期：2006-06-22)