文章编号: 1000-5013(2009)05-0517-03

一种实时说话人身份识别系统的设计

杨毅明, 陈东华

(华侨大学 信息科学与工程学院, 福建 泉州 362021)

摘要: 设计一种以TMS320VC5402为核心,外围扩展语音编解码器、程序存储器、语音存储器等集成电路,基于定点数字信号处理器的实时说话人身份识别的系统. 通过为每个语音命令设置3个模板的预留空间,使每个语音命令可以有3个模板参与识别,避免说话人语气和语调的变化产生的影响. 在比较嘈杂的环境下,系统的正确识别率超过87%. 处理器VC5402从识别中断并响应中断到CPU进入中断服务程序的第1条指令,需要20个时钟周期,系统的平均处理时间小于0.2 s.

关键词: 说话人; 身份识别; TMS320VC5402芯片; 数字信号处理系统

中图分类号: TN912.34 文献标识码: A

说话人身份识别是语音识别的特殊形式. 它和语音识别一样,都是事先对特定的语音信号进行处理,提取相应的特征或建立相应的模型;然后,根据这些特征或模型相对于其他语音信号做出判别. 由于说话人,说话速度、内容、环境的不同,语音信号具有多变性、动态性、瞬时性和连续性的特点. 识别这种语音信号的数字系统,需要有提取语音特征、建立语音模型、比较判断语音和输出识别结果等4个部分[1]. 本文介绍一种基于定点数字信号处理器的实时说话人身份识别的系统.

1 硬件的构成

系统硬件以高速的数字信号处理器为核心,结合语音编解码器、程序存储器、语音存储器、控制键盘等外围设备. 数字信号处理(DSP)和模拟信号的接口原理图,如图1所示.

数字信号处理器采用低功耗16 bit的TMS320VC5402芯片(简称C5402,美国TI公司),内部采用改良的哈佛结构. 优化的结构设计使其支持流水线操作,能够在单指令周期内完成乘法累加(MAC)运算,以及在单周期内执行3个操作数的指令,达到100 Mbit·s^{-1}指令的运算速度,指令周期为10 ns. C5402芯片的16 kB双寻址RAM可以保证系统算法程序在片内实时运行,1 MB的程序扩展空间可以保存算法处理的中间数据[2].

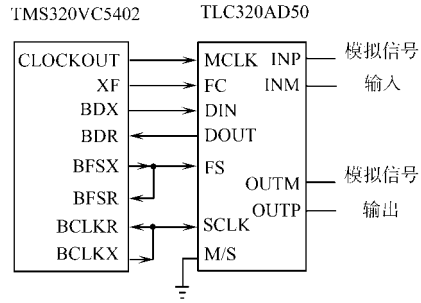

图1 接口原理图

Fig. 1 The principle of interface

语音编解码器采用高性能模拟接口电路TLC320AD50(简称AD50,美国TI公司),提供高分辨率的模拟与数字的信号转换. 该芯片由一对16 bit分辨率的同步串行变换通道组成,包括模数转换后的抽取滤波器和数模转换前的内插滤波器. AD50还有可编程2.00~22.05 kHz采样频率控制、可编程增益控制、帧同步延时、锁相环控制、通信协议等功能. AD50的典型模数转换和数模转换的信噪比是89 dB,动态范围是88 dB,它的数字接口直接与C5402的两个多通道缓冲串口McBSP之一连接(图1). 其中,C5402是主动方式,AD50是从动方式.

程序存储器采用型号M27C512的EPROM芯片,存储量是512 kbit,使能端$\overline{E}$控制电源和选择芯片. 当使能端$\overline{E}$高电平时,M27C512的电流从30 mA的读取方式降到100 μA的待机方式. 这时,输

收稿日期: 2008-01-12

通信作者: 杨毅明(1957-),男,教授,主要从事数字信号处理和信息编码解码的研究. E-mail: xululing@hqu.edu.cn.

出端是高阻抗状态. M27C512 的数据输出由输出使能端 $\overline{G}$ 控制, 在读取信号状态采用 5 V 电源. 当程序系统复位后, 语音程序自动由 EPROM 中加载进入 DSP 内部的程序存储区.

数据存储器采用型号 AT29C020 的 PEROM 芯片, 存储语音模板和信号处理的计算结果. 它的存储器分成 1 024 个区域, 每个区域能存 256 B, 它的工作电流和待机电流分别是 40 mA 和 100 μA, 每个存储区能反复地写入 1 万次以上, 读写数据都采用 5 V 电源. 当读取数据时, AT29C020 的$\overline{CE}$和$\overline{OE}$引脚是低电平、$\overline{WE}$引脚是高电平; 当写入数据时, AT29C020 的引脚是高电平$\overline{WE}$或引脚是低电平.

语音存储器采用 AT29C020 的随机存储器芯片, 键盘输入采用 74HC574 逻辑电路. 语音识别的系统采用单一的 5 V 电源, 其中 C5402 的 I/O 电路采用 3.3 V 电源供电、CPU 电路采用 1.8 V 电源供电, 属于双电源供电, 需要电源转换芯片. 它的稳压电源选用一个 TPS7301 和一个 TPS7333 电源转换芯片. 前者给 DSP 的 CPU 电路提供 1.8 V 的稳定电源, 后者给 DSP 的 I/O 电路提供 3.3 V 的稳定电源. 此外, 整个系统的供电采用微处理器监控电路 MAX705. 其优点是显著改善系统的可靠性和精确度, 降低电路的复杂性和减少元件的数量.

2 软件设计

处理器 C5402 有两个高速全双工 McBSP, 每个 McBSP 有 6 个引脚. 其中, 负责数据的是数据接收 DR, 数据发送 DX, 而负责同步的是移位发送时钟 CLKX、移位接收时钟 CLKR、帧同步发送时钟 FSX 和帧同步接收时钟 FSR. 为了使 C5402 与 AD50 进行正确通信, 必须对 C5402 的内部寄存器、McBSP 和 AD50 进行初始化, 使得 AD50 在采样频率 8 kHz 工作时, 每帧一个 16 bit 字的串行方式[3].

在实际环境中, AD50 获得的声音信号 $x(n)$ 包含环境声音和人说话的语音, 且没有语音的时间多于有语音的时间. 因此, 必须对 $x(n)$ 进行端点检测才能得到语音信号. 语音识别首先需要获得参考语音或参考模板, 然后, 将其他语音信号 $x(n)$ 与参考模板进行比较, 判断说话人的身份. 检测声音 $x(n)$ 的短时平均幅度和短时平均过零数可以确定语音的位置[4-5]. 选择空的长度为 50 点, 短时平均幅度为

$$M(n) = \sum_{m=-\infty}^{\infty} | x(m) | w(n-m), \tag{1}$$

其中, $n = 0, 1T, 2T, \cdots; T$ 是帧移; $w(n)$ 是截取一段声音信号的有限长窗序列. 这里, 选择窗的长度为 80 点, 则短时平均过零数为

$$Z(n) = \sum_{m=-\infty}^{\infty} | \operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)] | w(n-m). \tag{2}$$

式(2)中, sgn[ ]是符号函数. 设计 $M(n)$ 和 $Z(n)$ 程序是用电压低门限和高门限消除噪声的干扰, 同时结合时间门限, 防止突发噪声的干扰, 找出真正的语音段[6]. 当采样率为 8 kHz, 窗口长度为 10 ms, 共 80 个样本时, 短时平均幅度以最大值的 10% 作为语音起始点, 而以 20% 作为语音终点. 在样本幅度值归一化的情况下, 实验得到的浊音平均幅度是 12, 清音的平均幅度是 37. 参考模板的语音特征参数应该具有区别明显、相对独立和方便计算的特点, 才能确保语音身份识别能准确地实时进行. 建立的模板是 Mel 频率倒谱系数(Mel Cepstrum), 采用 Mel 刻度滤波器和离散余弦变换计算[7]. Mel 倒谱为

$$E_{\mathrm{mel}}(n, m) = \frac{1}{L}\sum_{l=0}^{L-1} \log[E_{\mathrm{mel}}(n, l)] \cos\left(\frac{2\pi}{L} lm\right). \tag{3}$$

式(3)中, L 是 Mel 滤波器组的三角滤波器个数; $E_{\mathrm{mel}}(n, l)$ 是 n 时刻的语音帧的第 l 个 Mel 刻度滤波器输出的能量. 即

$$E_{\mathrm{mel}}(n, l) = \frac{1}{A_l}\sum_{k=L_l}^{U_l} | V_l(\omega_k) X(n, \omega_k) |^2. \tag{4}$$

式(4)中, A_l 是三角滤波器的归一化因子, L_l 和 U_l 是三角滤波器的频率下限和上限, $V_l(\omega_k)$ 是三角滤波器的频率响应, $X(n, \omega_k)$ 是 $x(n)$ 的离散傅里叶变换. 在采样频率是 8 kHz、采样精度是 16 bit 时, 取 20 个三角滤波器、帧长为 256 点和帧移为 128 点建立模板.

模式识别或模式匹配的方法很多, 在孤立词说话人的身份识别中, 态时间规整(DTW)是最简单有效的方法. DTW 的本质是求测试模板 $\boldsymbol{T}$ 和参考模板 $\boldsymbol{R}$ 之间的最佳路径的积累距离 $D(n_N, m_M)$, N 是测试模板的语音帧总数, M 是参考模板的语音帧总数. 积累距离为

$$D(n_i, m_i) = d[\boldsymbol{T}(n_i), \boldsymbol{R}(m_i)] + D(n_{i-1}, m_{i-1}). \quad (5)$$

式(5)中, $d[\boldsymbol{T}(n_i), \boldsymbol{R}(m_i)]$是匹配点$(n_i, m_i)$上 $\boldsymbol{T}$ 的第n_i 帧语音特征矢量$\boldsymbol{T}(n_i)$, 以及 $\boldsymbol{R}$ 的第m_i 帧语音特征矢量 $\boldsymbol{R}(m_i)$的距离; $D(n_{i-1}, m_{i-1})$ 是前一匹配点(n_{i-1}, m_{i-1}) 的累积距离. 积累距离的计算是从第(n_1, m_1)点出发到达第(n_N, m_N)点, 得到总的积累距离.

设有j 个参考模板, 即有j 个待识别的说话人, 先根据式(5) 计算第 j 个人总的积累距离D_j; 然后, 求出$\{D_1, D_2, \cdots, D_j\}$, 最终的识别结果是$D_j$ 中最小者所对应的那个j. 将其与距离阈值进行比较, 小于阈值, 则j 为对应的说话人. 实验存储 3 个模版, 即保存 3 个说话人的语音特征.

3　讨论

实际应用的 DSP 源程序可以采用汇编语言或 C/C++ 语言编写, 关键的 DSP 程序一般还是用汇编语言编写. 这主要是因为大多数广泛使用的高级语言(如 C 语言) 并不适合描述典型的 DSP 算法. 典型的 DSP 应用都有大量的计算, 并有严格的开销限制, 使得程序的优化必不可少. 其次, DSP 的多存储器空间、多总线、不规则指令集、高度专业化的硬件等结构特点, 使得难以成为高级语言编写高效率的编译器. 再次, 对于底层硬件的控制, 用汇编语言编写和调试将更加地直观和高效. 系统为每个语音命令设置 3 个模板的预留空间, 每个语音命令可以有 3 个模板参与识别, 识别结果取其中的最接近者或平均值最近者. 这样, 可以避免说话人语气和语调的变化产生的影响, 提高 DTW 算法的一次识别率.

经过大量实验表明, 在比较嘈杂的环境下的正确识别率超过 87%. 处理器 C5402 从识别中断并响应中断, 到 CPU 进入中断服务程序的第 1 条指令需要 20 个时钟周期, 系统的平均处理时间小于 0.2 s. 时间较长是因为程序较多地使用 C 语言编程. 本系统经过调试, 证明总体设计思路正确, 方案可行, 满足性能指标要求. 另外, 对本系统编写不同的程序, 可以实现对不同的信号的采集和处理.

参考文献:

[1] 曾日波. 小词表实时语音识别系统的定点 DSP 实现[J]. 现代电子技术, 2004(11): 62-64.
[2] 尹　勇, 欧光军. DSP 集成开发环境 CCS 开发指南[M]. 北京: 北京航空航天大学出版社, 2003.
[3] 乔瑞萍, 崔　涛, 张芳娟. TM S320C54x DSP 原理及应用[M]. 西安: 西安电子科技大学出版社, 2005.
[4] 易克初, 田　斌, 付　强. 语音信号处理[M]. 北京: 国防工业出版社, 2000.
[5] 胡　航. 语音信号处理[M]. 哈尔滨: 哈尔滨工业大学出版社, 2000.
[6] 赵　静, 罗兴国, 蔡文涛. 噪声环境下语音信号的基音检测[J]. 语音技术, 2007, 31(3): 54-62.
[7] THOMAS F Q. 离散时间语音信号处理——原理与应用[M]. 赵胜辉, 等译. 北京: 电子工业出版社, 2004.

Design of a Real-Time Speakers Recognition System Based on TM S320VC5402

YANG Yi-ming, CHEN Dong-hua

(College of Information Science and Engineering, Huaqiao University, Quanzhou 362021, China)

Abstract: Designing a real-time speaker recognition system that is based on the fixed-point digital signal processor. The system is adopted the TMS320VC5402 as the handle center, its peripheries are speech codec, program memory, speech memory, etc. integrated circuits. By means of setting 3 templates reserved space for each speech command, making every speech command may have 3 templates participate identification, which can avoid the influence of speaker's tone and intonation changing. In fairly noise conditions, the system can have correct recognition over 87%. The processor of VC5402 needs 20 clock periods from recognize and respond interrupt to CPU enter interrupt serve program, the average time of system processing is less than 0.2 second.

Keywords: speaker; recognition; TM S320VC5402 chip; digital signal processing system

(责任编辑: 陈志贤　　英文审校: 吴逢铁)