基于语音识别的智能家居系统研究*

刘其洪, 李仲阳, 徐孟龙
(湖南师范大学 工学院，湖南 长沙 410081)

摘 要：针对智能家居系统中信息不能随时随地进行控制及其交流方式不流畅的问题，融合电话公用网和家庭网络设计了一个基于ARM9和语音识别技术的智能家居系统。该系统采用RASTA滤波方法去除语音信号中夹杂的卷积信道噪声，采用改进的动态时间规正（DTW）算法对语音命令进行识别。

关键词：智能家居；语音识别；带通滤波器；动态时间规正
中图分类号：TP391 **文献标识码**：A

Research of smart home system based on speech recognition

LIU Qi Hong, LI Zhong Yang, XU Meng Long
(College of Polytechnic, Hunan Normal University, Changsha 410081, China)

Abstract: A smart home system based on ARM9 and speech recognition technology was designed, which merged public telephone network and home network, and made smart home system be controlled at any time and any where and the communicating way be natural and smooth. In the system, RelAtive SepeTrAl(RASTA) filter was used to reduce the convolution channel noise mixed in speech signal, and the improved dynamic time warping (DTW) algorithm was adopted for recognizing speech command.

Key words: smart home; speech recognition; RASTA; dynamic time warping

随着计算机网络技术的发展，“智能家居”越来越被人们所重视。综观国内外的智能家居系统，大部分侧重于利用Internet进行远程控制。由于受到上网设备的限制，这种方式给智能家居系统的使用带来了不便。例如，在用户回家途中希望能够打开空调，很可能就因为不方便上网而无法实现。随着电话的普及，利用电话对家电进行远程控制可以做到随时随地。目前国内外使用的控制方式主要有：利用短消息控制和用语音播放受控设备的名称或代号，再根据用户的选择来控制相应的设备。这两种方式的弊端是：交流的方式不流畅自然，而且当受控设备的个数达到10个以上时需要考虑新出现的问题。

本文所介绍的系统采用三星公司生产的S3C2410芯片作为微处理器，基于公用电话交换网设计而成。系统先对用户从电话输入的语音命令进行语音识别，然后根据识别结果向串口发送相应的命令，实现对家电的控制。

1 语音识别原理

语音识别的过程可归结为模式识别和匹配。通过对语音信号进行预处理和分析计算可抽取出所需的语音特征，并以此建立语音识别所需的模板。而当对语音进行识别时，则需要将系统中存放的语音模板与输入的语音信号的特征进行比较，并根据一定的算法和策略，找出一系列最优的与输入的语音匹配的模板，最后输出识别结果，识别流程如图1所示[1]。

预处理包括采样、去除噪音、端点检测、预加重、分帧、加窗等。而特征参数的提取,目前较为常用的有线性预测倒谱系数(LPCC)与Mel倒谱系数(MFCC)。本文采用的是MFCC。系统采用改进的动态时间规正DTW

*基金项目：湖南省教育厅科研资助项目（08B050）

(Dynamic Time Warping)算法对语音进行训练和识别。

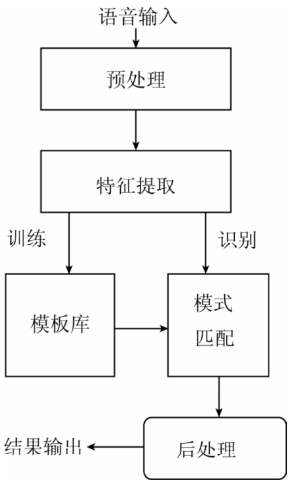

图1 语音识别过程

2 系统的硬件设计

2.1 系统的组成

系统由ARM9核微处理器构成主控部分，接收外部的控制信号，负责信息的处理，调控系统各部件协同进行工作。接口电路提供微处理器与电话网络的接口，包括振铃检测电路、模拟摘挂机电路、语音电路、双音多频（DTMF）信号解码电路。系统的硬件框图如图2所示。其中振铃检测电路、模拟摘挂机电路可由电话接口模块替代。

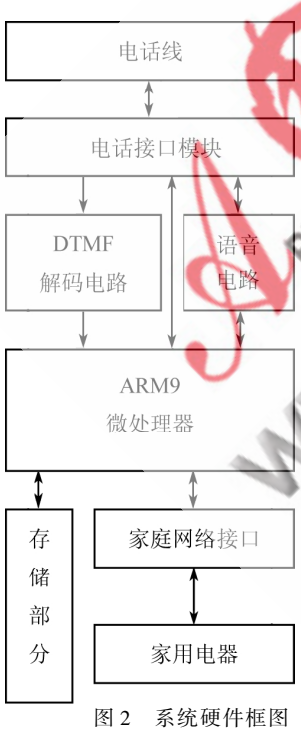

图2 系统硬件框图

系统的工作过程是：当有电话接入时，电话接口模块中的振铃检测电路检测到有振铃信号并送微处理器；微处理器发送摘机信号到摘挂机电路，实现模拟摘机；语音模块立即播放语音提示，要启动家电控制则输入密码，正常通话则等待3s；当用户有按键输入时，相应的DTMF信号经双音多频信号解码电路解码并送微处理器对它进行判断；若输入的密码正确，则由语音模块播放提示音，要求用户输入语音控制命令；语音命令的模拟信号经A/D转换电路转换成数字信号后送微处理器进行语音识别；识别结果通过串口输出。

2.2 硬件的设计

系统以S3C2410芯片为核心，配合电话接口模块、语音单元、存储单元实现语音识别的训练及识别过程。该芯片是SAMSUNG公司生产的一款基于ARM920T内核的芯片，拥有独立的16 KB指令Cache和16 KB数据Cache，片内资源丰富并支持MMU虚拟内存管理单元，采用它作为程序的主控芯片可减少工作量，降低开发难度。为实现对语音的处理，系统的语音单元采用了PHILIPS公司的UDA1341TS语音编解码芯片。语音的模拟信号从电话线经过偏置和滤波处理后输入到UDA1341TS中转换为数字信号后再送到S3C2410进行处理。本系统和电话公用网之间的接口是PH8809电话模块，PH8809是专业设计的电话接口电路，采用标准DIP32P封装，体积小巧，具备振铃检测、摘挂机检测和控制及语音接收/输出、DTMF输入/输出等功能，并带有电话线断线检测端口及音量自动增益调节电路，集成度高，性能稳定。出于安全考虑，系统采用密码验证的方式对用户身份进行识别，而密码的输入借助于DTMF信号。HT9170是DTMF信号接收解码芯片，它可对接收到的DTMF信号进行检测和解码，并将不同的DTMF输入信号转换成相应的4位BCD码数字信号输出。系统的存储部分选用32 M × 8 bit Flash K9F5608U芯片，用于烧写程序，2片8 M × 32 bit大容量SDRAM芯片型号为HY57V561620。其中电话接口模块PH8809与语音编解码芯片UDA1341TS及DTMF解码电路HT9170、微处理器S3C2410的连接如图3所示。

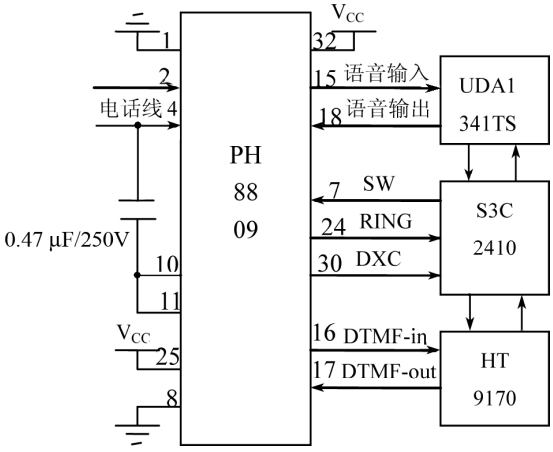

图3 电话接口模块与各电路的连接

图3中，SW是摘挂机控制开关，高电平导通，低电平断开。RING为振铃信号输出，输出高电平时无振铃，输出低电平则表示有振铃。DXC是电话线断线检测输出，输出高电平表示电话线断线，输出低电平时电话线连接正常。

3 系统的软件设计

系统首先通过移植vivi BootLoader、Linux操作系统建立系统的开发环境，然后再开发语音识别程序及硬件驱动程序并把它们烧写进目标板。其中重点难点在于语音识别程序的开发，本文只介绍此部分。

3.1 通信信道噪声的消除

基于电话的语音识别不同于普通的桌面语音识别，要想达到较好的识别效果，噪声的影响不能忽略。而RASTA滤波处理正是通过一个低端截止频率很低的带通滤波器对语音参数的时间轨迹进行滤波处理，以使频谱中的常量或者缓慢变化的部分得到抑制。系统引入了RASTA滤波技术并把它应用到Mel对数谱上，使得变化缓慢的通道噪声得到抑制[2]。

MFCC通过构造人的听觉模型，以语音通过该模型（滤波器组）的输出为声学特征，经过离散傅里叶变换（DFT）后，可得MFCC为

$$C(n)=\sum_{k=1}^{M}f(k)\cos[\pi(k-0.5)n/M] \tag{1}$$

式(1)中，$f(k)$为第k个滤波器的对数输出，n为MFCC的阶数，M为滤波器的个数。

设RASTA滤波器的系统函数为$H(z)$，则

$$H(z)=G\times\frac{Z^{N-1}\sum_{n=0}^{N-1}\left(\frac{N-1}{2}-n\right)Z^{-n}}{1-\rho Z^{-1}} \tag{2}$$

通常取$N=5$，$G=0.1$，$\rho=0.98$，这时

$$H(z)=0.1\times\frac{Z^{4}(2+Z^{-1}-Z^{-3}-2Z^{-4})}{1-0.98Z^{-1}} \tag{3}$$

又设$Y(k)$和$\overline{Y(k)}$分别代表RASTA处理前和处理后的第k个Mel频带对数频谱，则有：

$$\overline{Y(k)}=H(z)\times Y(k) \tag{4}$$

再对Mel频率对数频谱进行离散余弦变换(DCT)，可得：

$$\overline{C(n)}=\sum_{k=1}^{M}\cos\left(\frac{\pi(k-0.5)n}{M}\right)\overline{Y(k)} \tag{5}$$

式(5)中$\overline{C(n)}$是经过RASTA处理后的n阶MFCC。将式(4)代入式(5)，可得

$$\begin{aligned}\overline{C(n)}&=\sum_{k=1}^{M}\cos\left(\frac{\pi(k-0.5)n}{M}\right)\times H(z)\times Y(k)\\&=H(z)\times\sum_{k=1}^{M}\cos\left(\frac{\pi(k-0.5)n}{M}\right)\times Y(k)\\&=H(z)\times C(n)\end{aligned} \tag{6}$$

从式(6)中可以看出，RASTA处理完全可以从对数频率谱扩展到倒谱，即先求出MFCC，然后再做带通滤波处理，从而减少计算代价。

3.2 语音识别算法

在对语音信号提取MFCC特征参数及RASTA滤波去噪以后，语音信号就转化成为一组组特征向量，而语音识别算法的作用就是将待识别的语音信号的特征向量同系统中已建立起来的特征向量模板进行比较，找出最优的匹配模板。目前，常用的语音识别算法有隐马尔可夫模型(HMM)算法、动态时间规正(DTW)算法和人工神经网络(ANN)算法。其中，DTW算法具有系统开销小、运算速度快、对孤立词和小词汇表的识别简单而有效等特点，非常适合嵌入式系统的研制，而改进的DTW算法进一步减小了对计算量和存储空间的需求，因而本系统选用它作为系统的识别算法。

DTW算法是利用动态规划的思想，将一个复杂的全局最优化问题化为许多局部最优化问题来处理，并自动寻找一条路径，使两个特征矢量之间的积累失真量最小，从而避免由于时长不同而可能引入的误差。

设参考模板共有M帧矢量，待测语音模板共有N帧矢量（一般$M\neq N$），则动态时间归正就是寻找一个时间归正函数$m=\omega(n)$，它将测试矢量的时间轴n非线性地映射到模板的时间轴m上并使得测试矢量和模板矢量各帧之间的距离测度的累积和最小，从而使得两矢量之间的匹配路径最小，这样就保证了待测模板与参考模板之间具有最大的声学相似特性。通常，规正函数$m=\omega(n)$被限制在一个平行四边形（设为ABCD）网格内，它的起点坐标是(1，1)，终点坐标为(N，M)，相邻两边的斜率分别为2和1/2，如图4所示。

即只需对位于平行四边形ABCD内的各点对应的帧匹配距离进行计算即可，然而传统的DTW算法却对整

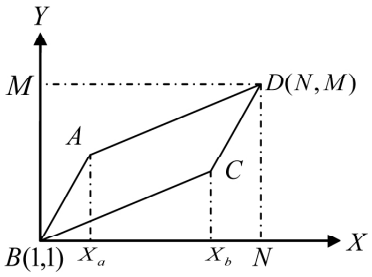

图4 匹配路径约束示意图

个矩形区域 MBND 都进行了计算，增加了系统的计算量。此外，传统的 DTW 算法还保存了所有的帧匹配距离矩阵和累积距离矩阵，而实际上每一列各个点上的匹配计算只用到了前一列的 3 个网格。改进的 DTW 算法对以上两点进行了改进，把实际的动态规正拆分为（1，Xa），（Xa+1，Xb），(Xb+1，N)3 段，其中，Xa 和 Xb 为最相近的整数且满足下式

$$\begin{cases} X_a = (2M - N)/3 \\ X_b = 2(2N - M)/3 \end{cases} \tag{7}$$

由此可以得出对 M 和 N 长度的限制条件

$$\begin{cases} 2M - N \geq 3 \\ 2N - M \geq 2 \end{cases} \tag{8}$$

当不满足以上条件时，认为两者差别实在太大，无法进行动态规正匹配。

而在 X 轴上的每一帧不再与 Y 轴上的每一帧进行比较，而只与 Y 轴上[y_{max}，y_{min}]间的帧进行比较，其中 y_{max}，y_{min} 由以下二式计算得到：

$$y_{\max} = \begin{cases} 2x & 0 \leq x \leq X_a \\ \frac{1}{2}x + (M - \frac{1}{2}N) & X_a < x \leq N \end{cases}$$

$$y_{\min} = \begin{cases} \frac{1}{2}x & 0 \leq x \leq X_b \\ 2x + (M - 2N) & X_b < x \leq N \end{cases}$$

当 $X_a > X_b$ 时，DTW 可拆分为(1，X_b)，(X_b+1，X_a)和(X_a+1，N)3 段，计算过程类似。

对于 X 轴上，每前进一帧，弯折特征都是一样的，累积距离的更新用下式实现

$$\boldsymbol{D}(x,y) = d(x,y) + \min[\boldsymbol{D}(x-1,y), \boldsymbol{D}(x-1,y-1), \boldsymbol{D}(x-1,y-2)]$$

上式中，矢量 $\boldsymbol{D}$ 用于保存前一列的累积距离，矢量 $\boldsymbol{d}$ 用于计算当前列的累积距离。根据上式，当在 X 轴上每前进一帧时，按上式可求出当前的累积距离，而它又可供下一列使用。如此不断的更新，当进行到待测模板的最后一帧时，矢量 $\boldsymbol{D}$ 的最后一个元素即为两个模板经过动态规正后的匹配距离。可以看出，该算法并没有像传统的 DTW 算法一样保存整个距离矩阵，从而节约了系统的存储空间[3]。

系统通过 DTMF 信号密码验证方式对用户身份进行识别，采用语音命令方式对家电进行控制，具有交流方式流畅自然、可实现随时随地控制、较高的安全性能等特点。实验结果表明，在一般的背景环境下，对孤立词的语音命令识别正确率达到 95% 以上，具有一定的应用价值。

参考文献

[1] 赵建光.嵌入式连续语音识别系统研究[D].河北工程大学硕士学位论文，2007.

[2] HERMANSKY H, MORGAN N. RASTA processing of speech[J]. IEEE Trans on Speech and Audio Processing, 1994, 2(4): 578–589.

[3] 林波,吕明.基于DTW改进算法的孤立词识别系统的仿真与分析[J].信息技术, 2006(4): 56–59.

（收稿日期：2009-02-24）

三剑侠持续升温，酷睿四核渐入主流

近日，由英特尔与中关村在线联手打造的“我的酷睿我超 Qiang 2009 英特尔酷睿英雄会”就在北京火爆开幕，这意味着英特尔的暑促正式开始。科技发烧友、DIY 高手、顶级超频玩家、电竞爱好者、英特尔精英俱乐部成员及网友代表聚集一堂，真是一场名副其实的英雄会。而英特尔暑促的核心产品英特尔三剑侠毫无疑问成为了主角。据悉，三剑侠也将随英雄会还将继续席卷南宁（7 月 6 日）、成都（7 月 10 日）及上海（7 月 16 日）。

或许有相当一部分消费者对“三剑侠”并不陌生，游戏铂金侠、高清黄金侠与网络白银侠，正如其名字中体现的那样，分别代表着在游戏、高清和网络等方面不同的应用模式。而这些基于英特尔酷睿架构的酷睿处理器家族，基本上能让你的电脑满足今天和明天的主流应用。随着价格逐渐平民化，四核处理器受关注的程度也越来越高。

从 Q8200 的测试报告里不难看出，基于酷睿?微体系架构的酷睿?2 四核处理器，其卓越的性能，使得视频编码和图片渲染速度也得到了极大的提升，轻松完成了动画制作的难题挑战，明显提高了动画设计师们的工作效率，而这也使得很多动画大片能够按期或者提前与观众见面。同时，其强大的处理能力，也给动画设计师们提供了宽阔的平台，最大限度地帮助他们实现心中的设计。