第33卷 第13期 计 算 机 工 程 2007年7月
Vol.33 No.13 Computer Engineering July 2007
·博士论文· 文章编号：1000—3428(2007)13—0025—03 文献标识码：A 中图分类号：TN912

基于片上系统的孤立词语音识别算法设计

刘金伟，黄樟钦，侯义斌
(北京工业大学计算机学院，北京 100022)

摘 要：介绍了孤立词语音识别系统，针对片上系统进行了语音识别算法的选择。对基于语音帧的端点检测算法、线性预测编码倒谱系数LPCC算法和动态时间规整DTW算法进行了分析和设计。对于新型语音识别SoC芯片的开发研制和推动片上可编程系统(SoPC)的研究与发展具有一定的理论和实践意义。

关键词：片上系统；语音识别；端点检测；LPCC；DTW

Design of Isolated Word Speech Recognition Algorithms for System on Chip

LIU Jinwei, HUANG Zhangqin, HOU Yibin
(College of Computer, Beijing University of Technology, Beijing 100022)

【Abstract】 This paper introduces the isolated word speech recognition for system on chip(SoC). The algorithms applied to isolated word speech recognition on SoC are chosen and designed. The paper analyzes and designs the endpoint detection algorithm, LPCC method and DTW method of speech frame. It is helpful towards the research and development on new types of speech recognition SoC and SoPC.

【Key words】 system on chip(SoC); speech recognition; endpoint detection; LPCC; DTW

目前，嵌入式语音识别系统主要通过单片机和DSP来实现。单片机运算速度慢、处理能力低，DSP虽然处理速度快，但是成本高、能耗大。因此，为了满足嵌入式系统的体积越来越小、功能越来越强的需求，语音识别片上系统SoC应运而生。语音识别SoC本身是一块芯片，在单一芯片上集成了模拟语音模数/数模转换器、信号采集转换器、处理器、存储器和I/O接口等，集成了声音信息的采集、取样、处理、分析和记忆。SoC有片内处理器和片内总线，具有速度快、体积小、成本低、可扩展性强等优点，已成为语音识别技术应用发展的一个重要方向[1]。研究和开发应用于片上系统SoC芯片的语音识别算法有着重要意义。

1 孤立词语音识别系统

孤立词语音识别系统应用于嵌入式控制领域，例如数字家庭控制、车载语音控制和智能语音可控玩具等。这种系统的原理如图1。

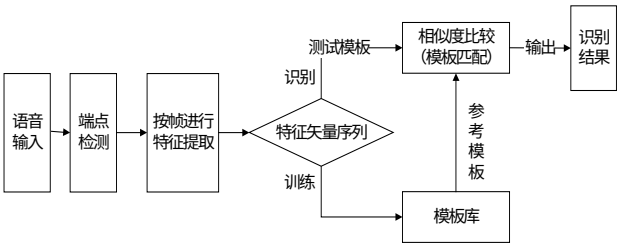

图1 孤立词识别系统原理

在训练阶段，用户将每个词依次说一遍，并将计算得到的每个词所对应的特征矢量序列作为模板存入模板库中。在识别阶段，将输入语音的特征矢量序列依次与模板库中的每一模板进行相似度比较，将相似度最高者作为识别结果输出。

2 针对SoC的孤立词语音识别算法设计

在SoC芯片中实现孤立词语音识别系统，就要根据片上系统的特点进行SoC语音识别算法的选择和设计。

(1)特征提取算法的选择。MFCC算法能很好地表征语音信号，而且在噪声环境下能取得很好的识别效果。而LPC系数对元音有较好的描述能力，对辅音描述能力较差，抗噪声性能也相对差些。但是考虑算法的计算量，MFCC提取特征参数是LPCC的10倍左右，通常在嵌入式系统下较难实现实时性，因此，选用LPCC算法。

(2)模式匹配技术的选择。隐马尔柯夫模型HMM方法是用概率及统计学理论对语音信号进行分析与处理的，适用于大词汇量、非特定人的语音识别系统。该算法对系统资源的要求较多。而动态时间规整技术DTW采用模板匹配法进行相似度计算，系统开销小、识别快，能有效节约系统资源、降低系统成本。由于嵌入式系统资源有限，语音命令识别系统所需要的词汇量有限，所须识别的语音都是简短命令，因此模式匹配算法选择DTW。

2.1 端点检测算法设计

一个好的端点检测算法可以在一定程度上提高系统的识别率。在双门限端点检测原理的基础上，进行语音端点检测算法的设计。为了提高端点检测的精度，采用短时能量E和

基金项目：国家自然科学基金资助项目(90407017)；北京市教委基金资助项目(KP2701200201)

作者简介：刘金伟(1972－)，男，工程师、博士研究生，主研方向：嵌入式系统，计算机网络等；黄樟钦，博士、教授；侯义斌，博士、教授、博士生导师

收稿日期：2006-07-28 **E-mail：**liujw@bjut.edu.cn

短时过零率 ZCR。

语音采样频率为 8KHz，量化精度为 16 位，数字PCM码首先经过预加重滤波器H(z)=2-0.95z^{-1}，再进行分帧和加窗处理，每帧 30ms，240 点为 1 帧，帧移为 80，窗函数采用Hamming窗。然后对每帧语音进行归一化处理，把值的范围从[-32767, 32767]转换到[-1, 1]。

实验发现，双门限端点检测算法对于 2 个汉字和 3 个汉字的语音命令端点检测效果不好。以语音“开灯”为例，如图 2 语音波形图中，端点检测只能检测到第 1 个字。

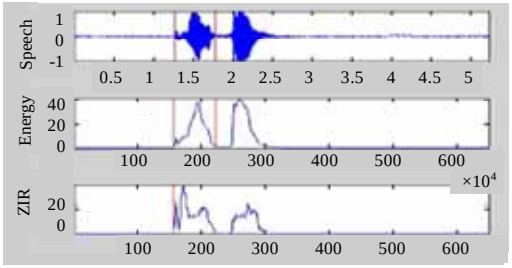

图 2　改进前对语音“开灯”的端点检测

如果语音命令中 2 个字的间隔过长，使用双门限端点检测法会发生只检测到第 1 个字的情况，从而可能造成语音匹配错误。

为避免该错误，把可容忍的静音区间扩大到 15 帧，如 15 帧内一直没有 energy 和 ZCR 超过最低门限，则认为语音结束；如发现仍然有语音，则算入在内。

改进后，整个语音信号的端点检测流程设计为 4 个阶段：静音，过渡段，语音段和语音结束。在静音段，如果能量或过零率超越低门限，就开始标记起始点，进入过渡段。在过渡段，由于参数的数值较小，不能确信是否处于语音段，因此只要 2 个参数的数值都回落到低门限以下，就将当前状态恢复到静音状态；而如果在过渡段中 2 个参数中的任何一个超过了高门限，就可以确信进入语音段。在语音段，如果两个参数的数值降低到低门限以下，且一直持续 15 帧，那么语音进入停止。如果 2 个参数的数值降低到低门限以下，但并没有持续到 15 帧，后续又有语音段越过低门限，那么认为语音还没有结束。如果检测出的这段语音总长度小于可接受的最小的语音帧数(设为 15 帧)，则认为是一段噪音而放弃。

采用改进后的端点检测算法，对单个汉字或多个汉字的语音命令均识别正常。图 3 为语音“开灯”的端点检测图(2 条竖线以内的部分为检测出来的语音部分)。

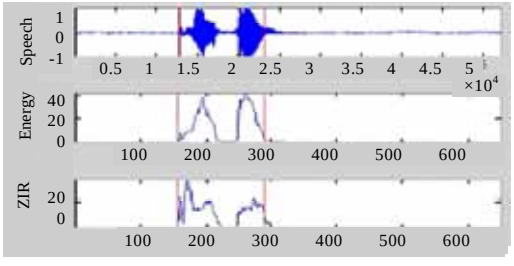

图 3　改进后对语音“开灯”的端点检测

2.2　LPCC 特征参数提取算法设计

LPC 参数是一种基于语音合成的特征参数。在应用中，使用的是由 LPC 系数推导出的线性预测倒谱系数(linear predictive ceptrum coefficients, LPCC)。

2.2.1 线性预测编码 LPC 算法

LPC模型的基本思想是：对于给定一n时刻采样的语音信号采样值$s(n)$，可以用p个取样值的加权和线性组合来表示[2]。$a_1, a_2,\ldots,a_p$称为LPC系数，也是全极点H(z)模型的　　参数。

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p) \tag{1}$$

对 LPC 语音数字模型来说，$u(n)$是输入，语音信号 $s(n)$是输出。数字模型的差分形式可表示 $s(n)$和 $u(n)$的时域关系为

$$s(n) = \sum_{i=1}^{p} a_i s(n-1) + Gu(n) \tag{2}$$

其中，$Gu(n)$是归一化冲击响应及其增益系数的乘积。该式的 Z 域表达形式为

$$s(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + Gu(z)$$

得到系统的传递函数：

$$H(z) = \frac{S(z)}{Gu(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \tag{3}$$

$H(z)$是声道模型和辐射模型的级联，是一个短时稳定的时变滤波器。语音信号的输出 $s(n)$可以用 p 个样本的线性组合表示，定义系统的预测输出为

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k z^{-k} \tag{4}$$

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k)$$

为计算 LPC 参数，定义起点为 n 的短时语音信号和误差信号。

$$s_n(m) = s(n+m) \tag{5}$$
$$e_n(m) = e(n+m)$$

误差平方和为

$$E_n = \sum_m e_n^2(m) = \sum_m | s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) |^2 \tag{6}$$

按照对预测误差的均方值最小的准则求a_k，则得到

$$\sum_m s_n(m-i) s_n(m) = \sum_{k=1}^{p} a_k \sum_m s_n(m-i) s_n(m-k) \tag{7}$$

根据相关函数定义，整理式(7)后得到

$$\phi_n(i,0) = \sum_{k=1}^{p} a_k \varphi_n(i,k) \qquad k = 1,2,\ldots,p \tag{8}$$

这是由 p 个方程构成的方程组，求解该方程组，得到系统的线性预测系数。系统的最小均方误差就可表示为

$$\hat{E}_n = \sum_m s_n^2(m) - \sum_{k=1}^{p} a_k \sum_m s_n(m) s_n(m-k) = \phi_n(0,0) - \sum_{k=1}^{p} a_k \phi_n(0,k) \tag{9}$$

用自相关法求解式(9)，求解采用高效的 Durbin 算法。

2.2.2　LPC 的倒谱计算 LPCC

实际中一般使用线性预测倒谱系数 LPCC，因为它能够较彻底地去掉语音产生过程中的激励信息，能较好地表征语音信号和反映声道响应，提高特征参数的稳定性。

倒谱定义为时间序列Z变换的模的对数逆变换。同态处理是将信号卷积运算转化为乘法运算，再取对数，将信号变成可以分离相加的形式。可推导出倒谱系数c_m与 LPC 的系数a_m之间的递推关系[3]：

$$\begin{aligned} c_0 &= \log G^2 \\ c_m &= a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} \qquad 1 < m < p \\ c_m &= \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} \qquad m > p \end{aligned} \tag{10}$$

其中，m 为倒谱系数的阶数；p 为线性预测系统的阶数。这样利用 LPC 系数通过递推公式可以得到 LPCC 参数。

LPCC 由于利用了线性预测中声道系统函数的最小相位

特性，避免了相位卷积、求复对数的复杂；且 LPC 倒谱的运算量小，仅是用 FFT 求倒谱时运算量的一半，适合于实时的语音识别片上系统。

2.3 基于 DTW 的模式匹配算法设计

2.3.1 动态时间规整 DTW 算法

存入模板库的各个词条称为参考模板，一个参考模板可表示为 $R=\{R(1), R(2),\ldots, R(m),\ldots, R(M)\}$。$m$ 为训练语音帧的时序标号，$m=1$ 为起点语音帧，$m=M$ 为终点语音帧，$R(m)$ 为第 m 帧的语音 LPC 倒谱特征矢量。所要识别的一个输入词条语音称为测试模板，可表示为 $T=\{T(1), T(2),\ldots, T(n),\ldots, T(N)\}$，$n$ 为测试语音帧标号，模式中共包含 N 帧语音，$T(n)$ 为第 n 帧的 LPC 倒谱特征矢量。

比较参考模板和测试模板的相似度，可以计算它们的距离 D[T,R]，距离越小则相似度越高。语音中各个段落在不同情况下的持续时间会产生长短的变化，大多数情况下测试模板和参考模板长度不相等，因此，这里采用动态规整的方法。

DTW 算法的实现为：分配 2 个 N×M 矩阵，分别为累积距离矩阵 $\boldsymbol{D}$ 和帧匹配距离矩阵 $\boldsymbol{d}$，其中，帧匹配距离矩阵 $d(i,j)$ 的值为测试模板的第 i 帧与参考模板的第 j 帧间的距离。算法分为两步：先计算参考模板的所有帧和未知模板的所有帧之间的相互距离，结果存在矩阵 $\boldsymbol{d}$ 中；然后根据判断函数计算累积距离矩阵 $\boldsymbol{D}$，$\boldsymbol{D}(N,M)$即为最佳匹配路径所对应的匹配距离。DTW 算法约束端点条件为：起点(1,1)，终点(n,m)，n 和 m 分别为测试和参考语音模板的帧序列长度。DTW 算法的计算流程如图 4。

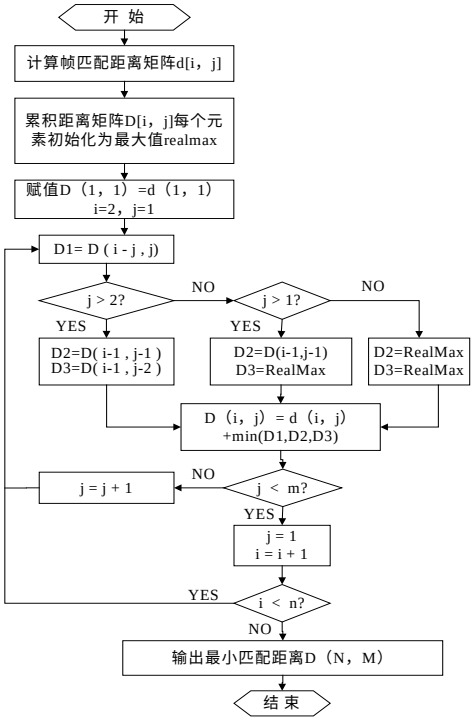

图 4 DTW 算法流程

2.3.2 应用中提高识别率的改进方法

应用中，提高DTW识别率的办法有冗余模板法、松弛起点终点法、改进局部路径约束函数法等[4]。本文采用的模板匹配算法就是融合这 3 种改进技术后的动态规整算法，称为ADTW算法。本文对这种算法进行了实验，得到该算法的实际识别率，并对之进行分析。

2.3.3 应用中提高算法效率的改进方法

在计算传统 DTW 算法过程中，如果限定动态规整的计算范围，就可以大大减小计算量，提高程序的性能。

采用平行四边形限制动态规整范围，如图 5，菱形之外的节点对应的帧匹配距离是不需要计算的，也没有必要保存所有帧匹配距离矩阵和累积距离矩阵，以ADTW的局部约束路径为例，每 1 列各节点上的匹配计算只用到了前两列的几个节点。这样可以减少计算量和存储空间的需求。把实际的动态规整分为 3 段：$(1,X_a)$，(X_a+1,X_b)和(X_b+1,N)。由于X轴上每前进一帧，只用到前两列的累积距离，因此只需要 3 个列矢量$\boldsymbol{A}$、$\boldsymbol{B}$和$\boldsymbol{C}$分别保存连续 3 列的累积距离。每前进一帧都对$\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$进行更新，即用$\boldsymbol{A}$和$\boldsymbol{B}$的值求出$\boldsymbol{C}$，再根据$\boldsymbol{B}$和$\boldsymbol{C}$的值求出下一列的累积矩阵放入$\boldsymbol{A}$中，由此可以反复利用这 3 个矢量，一直前进到X轴上最后一列，最后一个求出矢量的第M个元素即为 2 个模板动态规整的匹配距离。

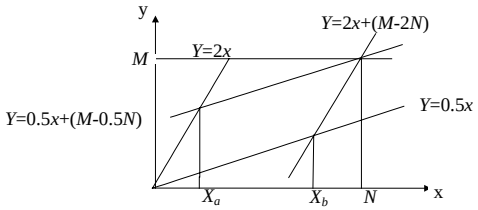

图 5 平行四边形区域限制

高效 DTW 对识别的区域进行了限制，整个平面区域大小为 $M\times N$。若 $M=N=150$，则限定区域内的计算量只是传统 DTW 算法计算量的 24%；同时由原来的 2 个 $N\times M$ 矩阵，减少为 3 个 M 矢量，大大减少了存储空间，解决了普通 DTW 算法在存储空间有限的嵌入式系统中难以实现的问题。

3 MATLAB 实验与分析

为了对算法的识别效果进行测试，本文设计了用于数字家庭控制系统的识别命令集，基于 Matlab 构建了孤立词语音识别系统，对识别算法进行实验和分析。该命令集包括语音命令 100 条，分别是：

(1)语音通话控制命令，如“拨打电话”、数字、人名；

(2)音乐控制类，如“打开音乐”、“减小音量”等；

(3)家庭电器控制类，如“打开空调”、“打开收音机”等。

该系统使用设计的端点检测技术、特征提取和模板匹配技术。语音采用频率为 8KHz，16 位量化精度，预加重系数 a=0.95，语音帧每帧 30ms，240 点为一帧，帧移为 80，窗函数采用 Hamming 窗。LPC 参数为 10 阶，LPCC 参数为 16 阶。实验人员为 3 名同学，两男一女，分别用甲乙丙代表，实验环境为办公室环境。实验平台为 WinXP，Matlab 7.0，进行的各项实验和数据分别如下：

(1)改进端点检测实验

针对设计的端点检测算法，得到改进前和改进后语音识别率的变化，数据见表 1。其中，改进前后所用的参考模板和测试模板数据相同。实验中 Matlab 程序自动把 50 个测试模板逐个和 100 个参考模板进行匹配，找到测试模板所对应的语音命令。

表 1 改进端点检测前后的识别率

	参考模板/个	测试模板/个	识别率/%	误识率/%
改进前	100	50	74	26
改进后	100	50	82	18

（下转第 55 页）