语音识别中 DTW 改进算法的研究

胡金平,陈若珠,李战明
(兰州理工大学 电气工程与信息工程学院,甘肃 兰州 730050)

摘 要:动态时间规整 DTW 是语音识别中的一种经典算法。对此算法提出了一种改进的端点检测算法,特征提取采用了 Mel 频率倒谱系数 MFCC,并采用计算量相对较小的改进的动态时间规整算法实现语音参数模板匹配,能够实现孤立词、特定人、小词汇量的语音识别,并用 Matlab 进行了算法仿真。试验结果表明,改进后的算法能够有效地提高系统对语音的识别率。

关键词:语音识别;端点检测;Mel 倒谱参数;动态时间规整
中图分类号:TP391.42 **文献标识码**:A **文章编号**:1674-7720(2011)03-0030-03

Discussion of improved DTW algorithm in speech recognition

Hu Jinping, Chen Ruozhu, Li Zhanming
(College of Electric and Information Engineering, Lanzhou University of Science and Technology, Lanzhou 730050, China)

Abstract: Dynamic time warping is a kind of classical programming in speech recognition. It adopts the improved endpoint detection algorithm and Mel frequence cepstrum coefficients to catch speech characteristic parameters and introduces dynamic time wrapping arithmetic to realize speech pattern matching. It is proved that this article designs a small vocabulary, isolated word speech recognition system, arithmetic of speech recognize simulate with Matlab software, the results show that the modified algorithm can provide a better performance in the speech recognition rate.

Key words: speech recognition; endpoint detection; MFCC; DTW

在孤立词语音识别中,最为简单有效的方法是采用动态时间规整 DTW(Dynamic Time Warping)算法,该算法基于动态规划(DP)的思想,解决了发音长短不一的模板匹配问题,是语音识别中出现较早、较为经典的一种算法。DTW 是把时间规整和距离测度计算结合起来的一种非线性规整技术,算法较为简洁,正确率也较高,在语音识别系统中有较广泛的应用。

本文对 DTW 算法提出了一种改进的端点检测算法,对提高系统的识别率有很好的实用价值[1]。

1 语音识别系统与 DTW 算法原理

本质上讲,语音识别就是语音信号模式识别[2],它由训练和识别两个过程完成。训练过程是从某一说话人大量语音信号中提取出该说话人的语音特征,并形成参考模式。识别过程是从待识语音中提取特征形成待识模式,与参考模式进行模式匹配、比较和判决,从而得出识别结果。本系统的结构如图 1 所示。

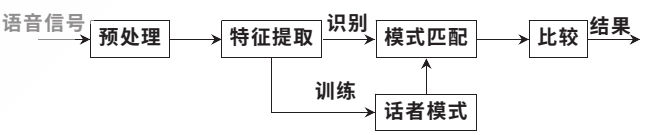

图 1 语音处理平台的系统图

假设测试和参考模板分别用 T 和 R 表示,它们之间的相似度用其之间的距离 $D[T,R]$ 来度量,距离越小相似度越高[3]。为了计算这一失真距离,要从 T、R 中各个对应帧之间的距离算起。设 n、m 分别是 T、R 中任意选择的帧号,$d[T(n),R(m)]$ 表示这两帧特征矢量之间的距离(在 DTW 算法中通常采用欧式距离)。

如图 2 所示,横轴上标出的是测试模板 T 的各个帧号 $n=1\sim N$,纵轴上是参考模板 R 的各个帧号 $m=1\sim M$,$N\neq M$。网格中的每一个交叉点 (n,m) 表示测试模式中某一帧与训练模式中某一帧的交汇点。DP 算法就是寻找一条通过此网格中若干个格点的路径。路径不是随意选择的,首先任何一种语音的发音快慢都有可能变化,但

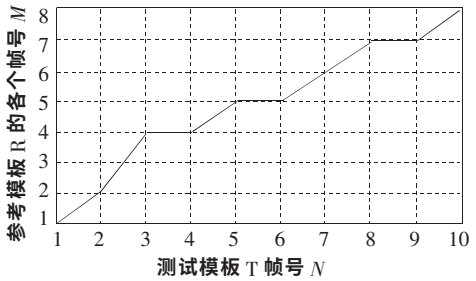

图 2 DTW 算法搜索路径

是其各部分的先后次序不可能改变，因此所选的路径必定是从左下角出发，在右上角结束。

假设路径通过的格点依次为：(n_1,m_1)，…，(n_i,m_i)，…，(n_N,m_M)，其中$(n_1,m_1)=(1,1)$，$(n_N,m_N)=(N,M)$。路径可以用函数 $m_i=\Phi(n_i)$来描述，其中 $n_i=i$，$i=1,2,\cdots,N$，$\Phi(1)=1$，$\Phi(N)=M$。为了使路径不至于过分倾斜，约束斜率设在 0.5~2 的范围内。如果路径已通过了格点(n_{i-1},m_{i-1})，那么下一个通过的格点(n_i,m_i)只可能是$(n_{i-1}+1,m_{i-1}+2)$、$(n_{i-1}+1,m_{i-1}+1)$和$(n_{i-1}+1,m_{i-1})$。用 η 表示这种约束条件，求最佳路径的问题则可以归结为满足约束条件 η 时，求最佳路径函数 $m_i=\hat{\Phi}(n_i)$，使得沿路径的积累距离达到最小，即：

$$\sum_{\substack{n_i=1\\ m_i=\Phi(n_i)\subset\eta}}^{N} D[n_i,m_i]=\min_{\Phi(\cdot)}\sum_{\substack{n=1\\ m_i=\Phi(n_i)\subset\eta}}^{N} D[n_i,m_i] \tag{1}$$

从$(n_1,m_1)=(1,1)$开始往下搜索(n_2,m_2)，再搜索(n_3,m_3)……，对每一个(n_i,m_i)都存储相应的前一格点(n_{i-1},m_{i-1})及相应的帧匹配距离 $d[n_i,m_i]$。搜索到(n_N,m_M)时，只保留一条最佳路径。由于 DTW 不断地计算测试矢量与模板矢量的距离以寻找最优的匹配路径，所以得到的两个矢量匹配是累计距离最小的路径函数，这保证了它们之间存在最大的声学相似特性。

2 语音识别改进算法的实现

2.1 语音信号的端点检测

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。输入的语音信号 $x(l)$，加窗分帧处理后得到的第 n 帧语音信号为 $x_n(m)$($w(m)$为汉明窗)，则：

$$x_n(m)=w(m)\times x_n(n+m)$$

其中，$m=0\sim N-1$(N 为帧长)；$n=0,T,2T$(T 为帧移)。

第 n 帧语音信号 $x_n(m)$的短时能量 E_n 为：

$$E_n=\sum_{m=0}^{N-1}|x_n(m)| \tag{2}$$

一帧信号中波形穿越零电平的次数称为过零率。定义 $x_n(m)$的短时过零率 Z_n 为：

$$Z_n=\frac{1}{2}\sum_{m=0}^{N-1}|\mathrm{sgn}[x_n(m)]-\mathrm{sgn}[x_n(m-1)]| \tag{3}$$

式中，sgn[ ]是符号函数。

为了提高端点检测的精度，采用短时能量和过零率进行端点检测[4]。语音采样频率为 8 kHz，量化精度为16 bit。数字 PCM 码首先经过预加重滤波器 $H(z)=1-0.95z^{-1}$，再进行分帧和加窗处理。在实验中发现，双门限端点检测算法对于两个汉字和三个汉字的语音命令端点检测效果不好。以语音“你好”为例，如图 3 语音波形图中，端点检测只能检测到第 1 个字。

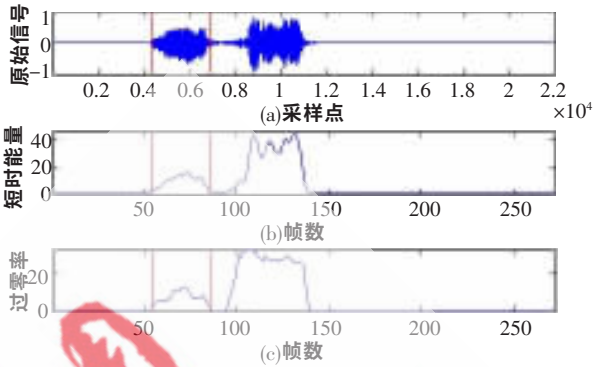

图 3 改进前语音“你好”的端点检测

如果语音命令中两个字的间隔过长，使用双门限端点检测法会发生只检测到第一个字的情况，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 帧内一直没有能量和过零率超过最低门限，则认为语音结束；如发现仍然有话音，则把能量和过零率计算在内[5]。

整个语音信号的端点检测流程设计为四个阶段：静音段、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，因此只要两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；而如果在过渡段中两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，则语音进入停止；如果两个参数的数值降低到低门限以下，但并没有持续到 15 帧，后续又有语音超越过低门限，则认为还没有结束；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 4 为语音“你好”的端点检测图。

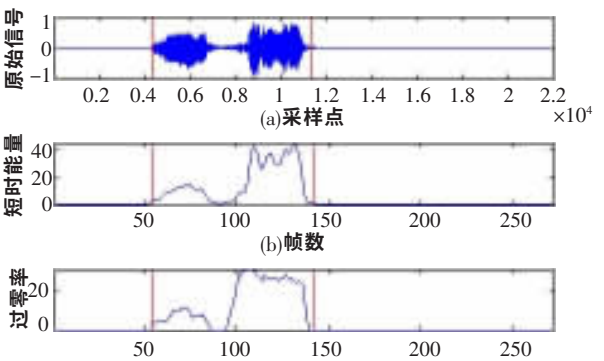

图 4 改进后语音“你好”的端点检测

2.2 语音识别的 DTW 高效算法

通常，路径函数 $\Phi(n_i)$被限制在一个平行四边形内，平行四边形的一条边斜率为 2，另一条边的斜率为 1/2。路径函数的起点为(1，1)，终止点为$(N，M)$。$\Phi(n_i)$的斜率为 0、1 或 2。这是一种简单的路径限制，如图 5 所示。

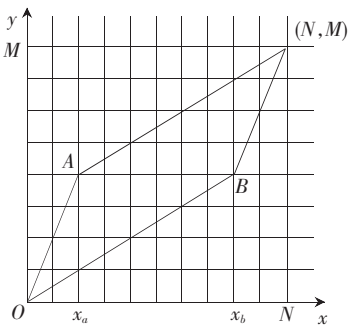

图 5 匹配路径约束示意图

本文的目的是寻找一个路径函数，在平行四边形内由点(1，1)到点$(N，M)$具有最小代价函数。由于对路径进行了限制，在匹配过程中许多格点实际上是到达不了的，因此，平行四边形之外的格点对应的帧匹配距离是不需要计算的。另外，也没有必要保存所有的帧匹配距离矩阵和累积距离矩阵，因为每一列各格点上的匹配计算只用到了前一列的 3 个网格。利用这两个特点可以减少计算量和存储空间的需求。

把实际的动态弯折分为三段：$(1，X_a)$、$(X_a+1，X_b)$和$(X_b+1，N)$，其中：

$$\begin{cases} X_a=\dfrac{1}{3}(2M-N) \\ X_b=\dfrac{2}{3}(2N-M) \end{cases} \tag{4}$$

由于 X_a、X_b 取最相近的整数，由此得出对 M、N 长度的限制条件：

$$\begin{cases} 2M-N\geqslant 3 \\ 2N-M\geqslant 2 \end{cases} \tag{5}$$

当不满足以上条件时，认为两者差别实在太大，无法进行动态弯曲匹配。在 X 轴上的每一帧不再需要与 Y 轴上的每一帧进行比较，而只是与 Y 轴上$[y_{\min}，y_{\max}]$间的帧进行比较。$y_{\min}$、$y_{\max}$ 的计算如下：

$$y_{\min}=\begin{cases} \dfrac{1}{2}x，0\leqslant x\leqslant X_b \\ 2x+(M-2N)，X_b<x\leqslant N \end{cases} \tag{6}$$

$$y_{\max}=\begin{cases} 2x，0\leqslant x\leqslant X_a \\ \dfrac{1}{2}x+(M-\dfrac{1}{2}N)，X_a<x\leqslant N \end{cases} \tag{7}$$

如果出现 $X_a>X_b$ 的情况，此时弯折匹配的三段为(1，X_b)、$(X_b+1，X_a)$和$(X_a+1，N)$。沿 X 轴上每前进一帧，虽然所要比较的 Y 轴上的帧数不同，但弯折特性是一样的，累积距离的更新都是用下式实现：

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

由于 X 轴上每前进一帧，只需要用到前一列的累积距离，所以只需要两个列矢量 D 和 d 分别保存前一列的累积距离和计算当前列的累积距离，而不用保存整个距离矩阵，这样可达到减少存储量和存储空间的目的。

2.3 试验结果

本系统采用改进的端点检测方法，采用 MFCC(Mel Frequene Cepstrum Coeffiients) 特征提取和 DTW 算法来实现语音识别。语音采样频率为 8 kHz，16 bit 量化精度，预加重系数 a=0.95，语音每帧为 30 ms，240 点为一帧，帧移为 80，窗函数采用 Hamming 窗。采集 5 个女生，10 个男生的数据。共分为两组，第一组是对 0～9 十个数字的识别，第二组是对孤立词的识别，试验数据如表 1 所示。

表 1 试验数据

算法	参考模板个数	测试模板个数	正确识别个数	错误识别个数	识别率/%
传统算法数字识别	150	150	137	13	91.3
改进算法数字识别	150	150	145	5	96.7
传统算法孤立词识别	150	150	127	23	85.7
改进算法孤立词识别	150	150	136	14	90.7

本文研究了语音识别 DTW 算法和理论，在应用中对双门限端点检测算法作了延长可容忍静音的改进，在说话语音识别算法上对 DTW 进行了改进和设计，实验结果表明，该算法可以有效地提高系统的识别率。

参考文献

[1] 何强，何英.MATLAB 扩展编程 [M].北京：清华大学出版社，2002.

[2] CHANWOO K，KWANG D S. Robust DTW -based recognition algorithm for hand -held consumer devices [J]. IEEE Transactions on Consumer Electronics，2005，51(2)：699-709.

[3] MIZUHARA Y，HAYASHI A，SUEMATSU N. Embedding of time series data by using dynamic time warping distances [J]. Systems and Computers in Japan，2006，37(3)：1-9.

[4] BDULLA A，CHOW W H，SIN D，G. Cross -words reference template for DTW -based speech recognition systems [C]. Conference on Convergent Technologies for the Asia-Pacific Region，TENCON，2003，2003：1576-1579.

[5] 刘金伟，黄樟钦，侯义斌.基于片上系统的孤立词语音识别算法设计[J]计算机工程，2007，33(13)：25-27.

(收稿日期：2010-09-04)

作者简介：

陈若珠，女，1963 年生，高级工程师，主要研究方向：语音识别，嵌入式。

胡金平，男，1985年生，硕士研究生，主要研究方向：语音识别，嵌入式研究。