文章编号:1006-9348(2009)10-0348-04

一种基于DTW的孤立词语音识别算法

张 军,李学斌
(北京化工大学信息科学与技术学院,北京 100029)

摘要:针对动态时间规整(DTW)对孤立词端点检测准确性过度依赖的问题,针对上述问题,采用放宽端点和限定动态规整计算范围结合的算法,不仅更准确的放松前后端点降低端点检测的敏感度,而且结合对动态规整计算范围的限定,减少计算量,提高执行效率。分别测试了基于传统DTW算法的识别率和改进后DTW算法的识别率。实验结果表明,改进后的算法,能有效提高孤立词识别率。

关键词:语音识别;动态时间规整;端点检测;孤立词
中图分类号:TN912.3 **文献标识码**:B

A Method of Isolated Word Recognition Based on DTW

ZHANG Jun, LI Xue-bin
(College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

ABSTRACT: The result of isolated word recognition of the DTW depends much on the accuracy of endpoint detection. In order to reduce the influence, a modified method of endpoint detection, for relaxing the endpoint of speech signals combining with the dynamic rule computing range restriction is proposed to overcome this disadvantage. It can not only more accurately relax the forward and back endpoint to reduce its sensitivity, but also reduce the amount of calculation and improve the performance with the dynamic rule computing range restriction. Compared the traditional DTW with modified DTW algorithm, the results show that the modified algorithm can provide a better performance in the speech recognition rate.

KEYWORDS: Speech recognition; DTW; Endpoint detection; Isolated word

1 引言

动态时间规整(DTW)是语音识别中把时间规整和距离测度计算结合起来的一种非线性归整技术[1],它利用动态时间伸缩算法有效地解决了孤立词识别时说话速率不均匀造成的时间伸缩问题,但对端点检测的精度依赖性较大,端点检测精度会随着不同的语音而有所不同,不能简单地将输入参数和相应的参考模板直接比较[2],因此要找到最佳匹配点需要考虑到多种可能的情况。

本文利用限定动态规整计算范围和松弛端点相结合的方法,从DTW路径搜索算法入手,规避一些不必要搜索,并缓和孤立词识别时对端点检测准确性过度依赖的问题。

2 孤立词识别系统

DTW算法是一种模板匹配的算法,设参考模板特征序列为:

$A=\{a_1,a_2,\cdots\cdots,a_I\}$,输入语音特征矢量序列为 $B=\{b_1,b_2,\cdots\cdots,b_J\}$,$I\neq J$。测试语音参数共有 N 帧矢量,参考模板共有 M 帧矢量,且 $N\neq M$。DTW算法的目的就是要找到一个最佳的时间规整函数 $J=W(I)$,并满足:

$$D=\min_{w(i)}\sum_{i=1}^{M}d[T(i),R(w(i))] \quad (1)$$

使得语音输入B的时间轴J映射到参考模板A的时间轴I上[3],总的累计失真最小。上式中 $d[T(i),R(\omega(i))]$ 是第I帧测试矢量 $T(i)$ 和第J帧模板矢量 $R(j)$ 之间的距离测度。D是最优情况下两矢量之间的匹配路径[4]。DTW采用逆向思路,从过程的最后阶段开始,逆推到起始点,寻找其中的最优路径。

基于DTW算法的孤立词识别系统主要由以下几个部分组成:语音输入、预处理、端点检测、语音特征参数提取、模板匹配和识别结果输出[5],如图1所示。

其中预处理主要包括预加重(为了加强语音高频的部分,补偿语音谱的固有衰落,消除唇辐射的影响)、分帧和加

收稿日期:2008-09-02 修回日期:2008-09-17

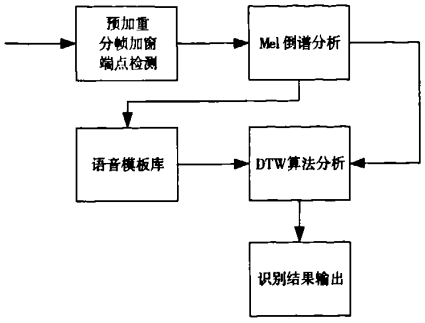

图1 孤立词语音识别系统结构

窗(为了消除各个帧两端可能会造成的信号不连续性)三个步骤。常用的预处理有数字滤波以及端点监测两种方法,本文使用端点检测来确定语音数据的起始点和终点。

3 松弛法提高检测端点的精度

在语音信号的预处理中,端点检测是关键的一步,语音信号的模型参数和噪声模型参数以及自适应滤波器中的适应参数都得依赖对应的信号段(语音段或噪声段)来计算确定[6]。因此,只有准确地判定语音信号的端点,才能正确地进行语音处理。

端点检测的目的是从包含语音的一般信号中确定出语音的起点以及终点,一般采用平均能量或平均幅度值与过零率相乘的方法来判断[7],对于传统的 DTW 算法,端点信息是作为一组独立的参数提供给识别算法,当环境噪声比较大或语音由摩擦音构成时,端点检测不易进行。利用松弛端点限制方法,将匹配过程中的固定起点(终点)改为松弛起点(终点),也就是路径不再是从(1,1)点出发,可以从(n,m)=(1,2)或(2,1),或(1,3)或(3,1)…点出发,称为松弛起点。同样,路径终点也不必在(N,M)点结束,可在(n,m)=(N,M-1)或(N-1,M)或(N,M-2)或(N-2,M)…点结束,称为松弛终点。一般情况下,起点和终点在纵横两个方向上放宽2-3帧。松弛起点终点的优点是克服由于端点检测不精确造成测试模板和参考模板起点终点不能严格对齐的问题,使DTW算法的搜索范围更具一般性。放宽端点前后的路径搜索范围如图2、图3所示。

针对距离测度计算误差,利用欧式距离公式来计算积累误差,

$$d = \sqrt{(a_0 - b_0)^2 + (a_1 - b_1)^2 + \cdots + (a_i - b_i)^2} \quad (2)$$

选用的衡量两帧(i 和 j)倒谱距离的标准就是这两帧倒谱的欧式距离。

$$D_{i,j} = D[(c_i, c_j)] = \|c_i, c_j\| = \sum_{k=1}^{Q} | c_i(k) - c_j(k) |^2 \quad (3)$$

路径通过网络点(n_0, m_0),(n_1, m_1),…(n_i, m_i)由(2)式

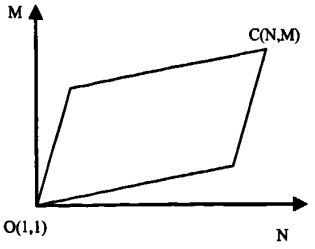

图2 固定起终点路径搜索范围

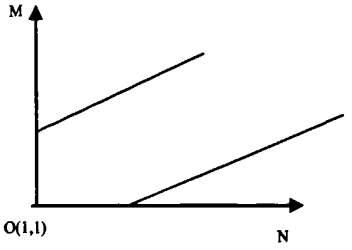

图3 松弛起终点后路径搜索范围

计算出路径的积累误差为

$$d[(n_{i-1}, m_i)] = D[T(n_{i-1}), R(m_{i-1})] + d[(n_{i-1}, m_{i-1})] \quad (4)$$

为了使识别更有效,将当前匹配点延不同路径递推到下一个匹配点赋予不同的权值,这样可以搜索到一条最佳路径,并为短时能量和短时过零率设置两个门限。一个是比较低的门限,对信号的变化比较敏感。另一个是比较高的门限,信号必须达到一定的强度,才被认为是语音开始。整个端点检测过程可以分为四段:静音、过渡段、语音段、结束[8]。图4为语音信号"0"采用传统算法的端点检测结果,图5为语音信号"0"采用改进算法后的端点检测结果。对比两图可以看出,采用改进算法后可以减少开始语音输入时算法对噪音的理出,端点检测结果比较准确。

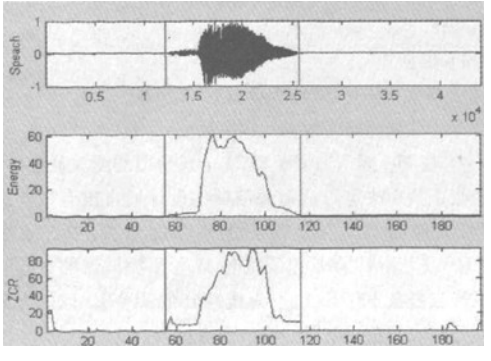

图4 传统端点检测算法的检测结果

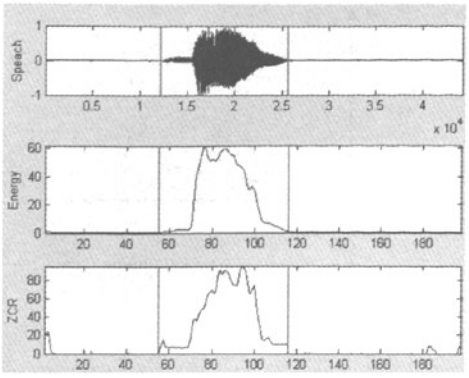

图5　改进端点检测算法的检测结果

4　Mel 倒谱分析

特征提取的目的是从语音信号中提取出对语音识别有用的信息，它对语音信号进行分析处理，去掉与语音识别无关的冗余信息，获得影响语音识别的重要信息。从信息论角度讲，这是信息压缩的过程。倒频谱有着能将频谱上的高低频分开的优点，它的实质就是将信号作适当的同态滤波，将信号中的卷积关系变为乘积关系，随之作对数处理使之化为可分离的相加成分[9]。

Mel 频率倒谱特征参数（MFCC）具有更好的鲁棒性，所含的信息量比其它参数多，而且考虑了人类发声与接收声音的特性，能较好的表现语音信号，因此，考虑 MFCC 作为特征参数具有一定的优势，本文中 MFCC 的计算流程如下：

1）对输入语音帧加 Hamming 窗（为了避免加矩形窗时对 MFCC 系数在端点造成的截取误差，采用 Hamming 窗函数进行加窗）

$$\tilde{x}_l(n) = x_l(n)w(n), \quad 0 \leqslant n \leqslant N-1. \tag{5}$$

其中：

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leqslant n \leqslant N-1. \tag{6}$$

2）加窗后进行快速傅利叶变换，将时域信号转化为频域信号；

3）根据式

$$Mel(f) = 2595\lg(1 + f/700)^{[10]} \tag{7}$$

将线性频标转换为 Mel 频率尺度；

4）在 Mel 频率轴上配置 24 个三角形的滤波器组（由信号的截止频率决定），这组滤波器的设计通常没有很严格的限制，大体与人耳接受声音的感知度相关，中心频率在 1000Hz 以上和 1000Hz 以下的各 12 个。滤波器的中心频率间隔特点是在 1000Hz 以下为线性分布，1000Hz 以上为等比数列分布。三角滤波器的输出则为：

$$Y_i = \sum_{k=F_{i-1}}^{F_i} \frac{k - F_{i-1}}{F_i - F_{i-1}} X_k + \sum_{k=F_i+1}^{F_{i+1}} \frac{F_{i+1} - k}{F_{i+1} - F_i} X_k \quad i = 1, 2, \ldots, 24 \tag{8}$$

其中 X_k 为频谱上第 k 个频谱点的能量，Y_i 为第 i 个滤波器的输出，F_i 为第 i 个滤波器的中心频率。

5）对所有滤波器输出做对数运算，再进行离散余弦变换（DCT）即得到 MFCC；

6）为体现语音的动态特性，在语音特征中加入了一阶差分倒谱，其计算方法如下式所示：

$$\Delta c_l(m) = \sum_{k=-2}^{2} k c_{l-k}(m) \quad 1 \leqslant m \leqslant P \tag{9}$$

其中下标 l 与 l－k 表示第 l 与 l－k 帧，m 表示第 m 维。

MFCC 参数计算的要点是将线性功率谱 S(n) 转换成为 Mel 频率下的功率谱，在计算前需要在语音频谱范围内设置若干带通滤波器 Hm(n)，m = 0..M－1，n = 0..N/2－1。M 为滤波器个数，N 为一帧语音信号的点数。每个滤波器具有三角形特性，中心频率为 fm，频率轴上是均匀分布的。在线性频率上，当 m 较小时，相邻的 fm 间隔很小，随着 m 的增加，相邻的 fm 间隔逐渐拉开。Mel 频率和线性频率的转换关系如下式：

$$mel = ln\left(1 + \frac{f}{700}\right) \cdot \frac{1000}{ln(1 + 1000/700)} \tag{10}$$

这组带通滤波器的参数是事先计算好的。

图 6 给出了 Mel 频率尺度滤波器组的分布图。

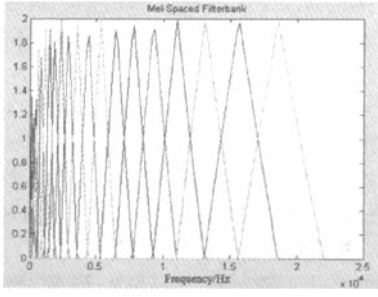

图6　Mel 尺度滤波器组

在对每一帧语音信号提取 MFCC 特征参数后，转化成了一组 MFCC 特征向量。本文中 DTW 语音识别就是要将测试语音的这个特征向量同模板库中已存在的语音特征向量进行模式匹配[11]，寻找距离最短的模式作为识别结果。

5　限制动态规整范围

松弛起终点后 DTW 路径搜索范围将变大，通过限定动态规整的计算范围，可以减小计算量，提高算法执行效率。本文在选取路径时，采用平行四边形限制动态规整范围，如图 7 所示。

在两个对应匹配帧各自时间轴所构成的矩形区域内，进行全面寻优要花费很多时间，这是语音识别系统不能允许的。考虑到说话人在说同一个字的时时间并不会相差太远，所以本文规定搜索的区域在斜率为 2 和 1/2 的两条斜线之

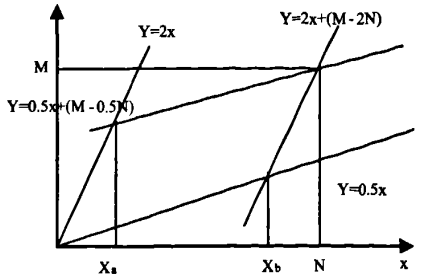

图7 平行四边形区域限制图

间,这样的搜索范围是有效的。平行四边形之外的节点对应的帧匹配距离不需要计算,也不用保存所有帧匹配距离矩阵和累积距离,实际的动态规整分为三段$(1,X_a)$,(X_a+1,X_b),(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,所以只需要三个列矢量A、B和C分别保存连续三列的累积距离,而不需保存整个距离矩阵。每前进一帧都对A、B、C进行更新,即用A和B的值求出C,再根据B和C的值求出下一列的累积矩阵放入A中,由此可以反复利用这三个矢量,一直进行到x轴上最后一列,最后一个求出矢量的第M个元素即为两个模板动态规整的匹配距离。

进行区域限制后,整个平面区域大小为M×N,匹配区域的大小为

$$space = \frac{8\sqrt{3}}{27}(M - \frac{1}{2}N)(2N - M) \qquad (11)$$

除了进行动态规整范围的限定,及时的删除不可能的点也是改进之一,在动态优化的过程中,每一步都会产生许多最小点的集合。在它们当中有一些已经超过了预先设定的阈值,所以可以将其舍弃,以节省大量的时间开销。

6 实验结果

基于以上算法,采用8kHz的采样频率,量化精度为8位,单声道采集0-9,对十个数字的语音进行识别。每个数字识别50次。录制的语音信号首先经过预加重滤波器

$$H(z) = 1 - \mu z^{-1} \qquad (12)$$

取μ为0.97。语音去噪结果如图8所示。

根据短时平稳特性,语音在10-30ms的范围内是稳定的,这里采用20ms,即160个数据为一帧,对帧数据进行MFCC提取。在训练阶段,将计算出的特征矢量序列作为模板存入模板库。在识别阶段,将输入语音的特征矢量系列与模板中的每一特征矢量序列进行模式匹配,匹配相似度最高者作为识别结果输出。由于发音长短的不同,即使同一个数字两次的发音可能有所区别,采用改进后的DTW算法进行识别,将采集的测试数据的特征与其他的样本数据进行比较,首先对正第一帧,然后将整体短的数据与长的数据进行动态对应,长的数据可以空1帧或者是直接对应,由于语音的伸缩性不会太大,所以最多空1帧,直到到达其中一个的末尾。

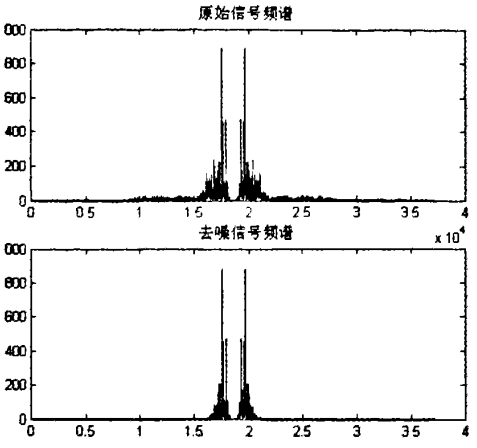

图8 语音去噪

实验过程中,将十个数字0-9分别统计,每个数字分两组各测试50次,即各取500个测试模板,把500个测试模板逐个与参考模板进行匹配,找到测试模板所对应的数字,在此基础上统计识别率,如表1和表2所示。

表1 传统DTW算法数字0-9的识别率

识别数	识别次数(个)	正确次数(个)	错误次数(个)	识别率
0	50	42	8	84%
1	50	44	6	88%
2	50	41	9	82%
3	50	42	8	84%
4	50	45	5	90%
5	50	44	6	88%
6	50	39	11	78%
7	50	46	4	92%
8	50	47	3	94%
9	50	41	9	82%
平均识别率		86.2%		

表2 改进后DTW算法数字0-9的识别率

识别数	识别次数(个)	正确次数(个)	错误次数(个)	识别率
0	50	47	3	94%
1	50	46	4	92%
2	50	46	4	92%
3	50	45	5	90%
4	50	46	4	92%
5	50	45	5	90%
6	50	41	9	82%
7	50	48	2	96%
8	50	48	2	96%
9	50	47	3	94%
平均识别率		91.8%		

(下转第364页)

参考文献：

[1] D C Chu. Phase Digitizing Sharpens Timing Measurements[J]. IEEE Spectrum, July 1988. 28 - 32.

[2] R F Schneider, D C Chu. Modulation Recognition Of Spread Spectrum Signals Using Modulation Domain Measurements[J]. MILCOM, IEEE 1991.

[3] 余翔. 调制域分析技术的研究[D]. 电子科技大学博士毕业论文, 1999 - 7.

[4] 陈俊. 调制域脉冲分析仪系统设计[D]. 电子科技大学硕士毕业论文, 2005 - 3.

[5] 李玉涛. 军用跳频通信综合测试仪——调制域分析模块的硬件系统设计[D]. 电子科技大学硕士毕业论文, 2005 - 5.

[6] 曲卫振,唐申生. 调制域测频原理及工程实现[J]. 仪器仪表学报, 1996,16(13).

[7] 曾纪瑞,秦开宇,曹勇. 调制域脉冲分析仪无死区触发设计[J]. 电测与仪表,2008 - 12.

[8] 曹勇,秦开宇. 一种调制域分析仪的测量原理与工程实现[J]. 仪器仪表学报增刊, 2007,28(4).

[9] 曹勇. 调制域分析仪系统设计及关键技术研究[D]. 电子科技大学硕士毕业论文,2007 - 5.

[作者简介]

曹　勇(1980.10 -),男(汉族),湖南邵阳人,博士生,电子科技大学助教,研究方向为导航与宽带无线通信测试技术、时频测量、调制域分析。

秦开宇(1967.8 -),男(汉族),四川遂宁人,工学博士,教授,博士生导师,院长,主要研究方向是导航与宽带无线通信测试、微波与通信测量技术。

(上接第 351 页)

基于传统 DTW 算法的孤立词平均识别率为 86.2%,采用改进后的 DTW 算法,正确识别率提高 5.6%,平均识别率达到 91.8%,识别结果优于传统 DTW 算法。

7　结束语

本文从 DTW 路径搜索算法入手,路径递推搜索时,采用松弛端点的办法,在选取路径时要求路径搜索范围在最大斜率为 2 和最小斜率为 1/2 的范围内,不断地计算矢量的距离以寻找最优的匹配距离路径,在得到的矢量匹配是积累距离最小的规整函数基础上,将模板特征序列和语音特征序列进行特征匹配,从而避免一些不必要的搜索,通过实验,进行孤立词(数字 0 - 9)的识别,在提高识别率上达到了更满意的识别效果。

参考文献：

[1] 严剑峰,付字卓. 一种新的基于信息熵的带噪语音端点检测方法[J]. 计算机仿真, 2005,22(11):117 - 119,139.

[2] 李大治,等. 基于在线垃圾模型的语音确认方法[J]. 计算机仿真, 2003,11(20):48 - 51.

[3] 王炳锡,屈丹,彭煊. 实用语音识别基础[M]. 北京:国防工业出版社, 2005.

[4] 张雄伟,陈亮,杨吉斌. 现代语音处理技术及应用[M]. 北京:机械工业出版社, 2003.

[5] 张钢,朱铮涛,何淑贤. 应用 DTW 的语音(声纹)鉴别技术研究[J]. 中国测试技术, 2007,2(33):120 - 123.

[6] 王让定,柴佩琪. 语音倒谱特征的研究[J]. 计算机工程, 2003,29(13):31 - 33.

[7] 刘鹏,王怀杰. 噪音环境下孤立词的语音识别[J]. 人工智能及识别技术, 2007,29(13): 1399 - 1404.

[8] 息晓静,等. 语音识别关键技术研究[J]. 计算机工程与应用, 2006,22(11):66 - 69.

[9] H Hermansky, H Morgan. RASTA processing of speech[J]. IEEE Trans on Speech and Audio Processing, 1994,2(4):578 - 589.

[10] HUANG Xuedong, A Acero, H W Hon. Spoken - Language - Processing[M]. Prentice Hall, 2001.

[11] Rabiner. Fundamentals of Speech Recognition [M]. Prentice Hall, 1992.

[作者简介]

张　军(1984 -),男(汉族),安徽六安人,硕士研究生,主要研究领域为数字信号处理及语音识别技术。

李学斌(1967 -),男(汉族),北京市人,副教授,主要研究领域为多媒体信息处理、滤波器组设计理论。