

一种新的基于 DTW 的孤立词语音识别算法

周炳良, 邓立新, 洪民江

(南京邮电大学 通信与信息工程学院, 江苏 南京 210003)

摘要: 为了提高传统孤立词语音识别系统中 DTW (动态时间规整) 算法的识别效率, 提出了一种新的基于 DTW 的孤立词语音识别算法。该算法针对语音识别阶段的方法进行改进: 首先对输入的测试语音信号进行预处理, 包括预加重、分帧加窗和端点检测, 然后提取预处理语音信号的特征参数矢量, 并通过截取测试语音特征矢量起始部分长度与库模板矢量进行最优路径匹配 (最优时间规整), 匹配后只保留失真度较小的部分库模板矢量继续进行下一轮最优路径匹配, 如此反复截取测试语音特征矢量的起始不同部分进行最优路径匹配与参考模板保留, 直至参考模板保留唯一。实验结果表明, 与传统 DTW 算法相比, 该算法在保证识别精度不变的前提下, 能大幅减少孤立词语音识别系统的计算开销, 有效提高孤立词语音识别系统的识别效率。

关键词: 语音识别; 动态时间规整; 部分匹配; 孤立词

中图分类号: TP301.6

文献标识码: A

文章编号: 1673-629X(2018)04-0119-05

doi: 10.3969/j.issn.1673-629X.2018.04.025

A Novel Isolated Word Algorithm of Speech Recognition Based on DTW

ZHOU Bing-liang, DENG Li-xin, HONG Min-jiang

(School of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In order to improve the recognition efficiency of dynamic time warping (DTW) algorithm in traditional isolated word recognition system, we propose a new isolated word speech recognition algorithm based on DTW. It has made improvements on the contents of speech recognition phase. Firstly, the input speech signal is preprocessed, including pre-emphasis, frame-by-frame and endpoint detection, and then the feature vector of preprocessed speech signal is extracted. The initial parts of speech feature vector are intercepted and respectively matched with all model vectors properly, then only several model vectors with less distortion can be left to continue next optimal path matching, repeated the optimal path matching and model vectors retaining until it is unique. The experiments show that compared with the ordinary DTW, the proposed algorithm can dramatically reduce the computational cost and improve the efficiency for the isolated word recognition system on the premise of ensuring its recognition accuracy.

Key words: speech recognition; DTW; partial match; isolated word

0 引言

语音识别即让机器接收、识别和理解语音信号, 能够“听懂”会话中的语音语义并执行人类意图。常用的识别方法包括动态时间规整 (DTW)、隐马尔可夫模型 (HMM) 和人工神经网络 (ANN) 等。在孤立词语音识别中, 动态时间规整是最简单有效的方法。DTW 算法基于动态规划 (DP) 的思想, 能够较好地解决孤立词识别时说话速度不均匀的难题。相较于传统的语音线性伸缩匹配的方法, DTW 方法有效提高了孤立词语音

识别系统的识别率, 因此在特定场合下得到了较好的应用。

近年来, 为了提高孤立词语音识别系统的效率, 使其广泛适用于市场和各类服务领域, 科研人员提出了许多新的基于 DTW 的语音识别算法。

文献[1]提出了基于音节个数的高效动态时间规整算法 (SEDW), 该算法利用彝语语音信号音节个数从 1 个到 7 个不等的特点, 预先检测出彝语语音信号中的音节个数, 并将其只与含有相同音节个数的模板

收稿日期: 2017-04-24

修回日期: 2017-08-29

网络出版时间: 2017-12-05

基金项目: 国家自然科学基金 (61401227)

作者简介: 周炳良 (1989-), 男, 硕士研究生, 研究方向为语音处理与人机交互; 邓立新, 副教授, 研究方向为现代语音信号处理。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20171205.0906.056.html>

进行最优匹配,减少了系统的计算开销,提高了系统的识别效率。但该算法利用双门限检测法分辨语音信号的各个音节,对门限阈值精度要求很高,一旦阈值设置不准确,系统识别效率将大幅降低,且该算法只适用于彝语语音信号识别,适用率较低。

文献[2]提出了改善局部路径限制的 DTW 算法,该算法改善了局部路径节点前进的范围,有利于解决测试语音特征矢量与模板矢量均匀变化剧烈的匹配问题,加快了两矢量匹配的过程。但该算法增加了系统局部路径搜索的复杂度和内存消耗,且不利于解决两矢量均匀变化平缓的匹配过程。

文献[3]提出了增设参考模板阈值的 DTW 算法,该算法在进行测试语音特征矢量与模板矢量匹配时,一旦计算出部分失真度大于预先增设的模板阈值,将终止对该模板继续运算,转入对其他模板继续匹配运算。由于是中途停止对模板的匹配运算,因此可以节省部分计算开销,提高了系统的识别效率。该算法的识别效率优于文献[2]的算法,但是算法必须要为每一个模板找到一个合理的阈值,否则将无法减少系统的运算量,甚至大幅度降低系统的识别率。

在识别阶段,传统的 DTW 算法在测试语音特征矢量与所有参考模板矢量之间进行的是全长度最优路径匹配,一旦系统的参考模板数量较大,系统的计算量和内存消耗将加剧,从而严重影响识别系统的效率。针对上述问题,提出了一种加快 DTW 模板匹配的改进算法:预先对提取得到的测试语音特征矢量进行部分长度截取,并将得到的部分特征矢量与模板矢量进行最优路径匹配,排除掉匹配度较小的部分模板。如此快速反复进行语音部分匹配和模板排除,直至模板数量唯一。通过实验对该算法的识别效率进行了验证。

1 DTW 算法的基本原理

DTW 技术是一种把时间规整和距离测度计算结合起来的非线性规整技术,通过不断地计算测试语音特征矢量和模板特征矢量的距离来搜索两者之间的最优时间规整(最优匹配路径),保证它们之间存在最大的声学相似特性。假设待测语音共有 N 帧矢量,参考模板共有 M 帧矢量,分别记为 T 和 R ,且 $N \neq M$,则动态时间规整就是要找一个时间规整函数 $m = w(n)$,将测试语音特征矢量的时间轴 n 非线性地映射到模板的时间轴 m 上,并使函数 w 满足式(1):

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

其中, $d[T(n), R(w(n))]$ 是第 n 帧矢量 $T(n)$ 和第 $m = w(n)$ 帧模板矢量 $R(m)$ 之间的距离测度; D 是

处于最优时间规整情况下两矢量的累积距离^[4]。

实际上,DTW 算法本质就是搜索测试语音特征矢量与模板矢量的最优匹配路径并求出两者之间最小累积距离 D 。现实应用中,由于说话人对同一语音的发音速率相差一般不超过 2 倍,为了满足这一实际特性,搜索最优匹配路径时,应该将最优匹配路径全局(任意节点与起点或止点的连线)限制在两边斜率分别为 $1/2$ 和 2 的平行四边形中^[5-6],如图 1 所示,且止点 (N, M) 满足式(2):

$$\begin{cases} 2N - M \geq 2 \\ 2M - N \geq 3 \end{cases} \quad (2)$$

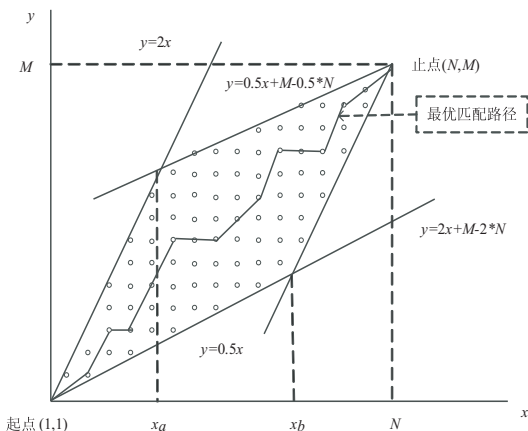


图 1 最优路径全局限制示意图

同时对其局部路径(任意节点与前续节点的连线)也应加以限制,确保搜索路径中节点的前续节点在一定范围内,典型的一种局部路径限制方式^[7-8]如图 2 所示。

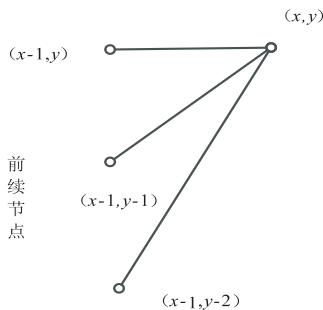


图 2 最优路径局部限制示意图

为了找到满足最小测度距离 D 的最优匹配路径,DTW 算法采用逆序决策过程,即节点的前续节点决定该节点的最小累积距离。假设路径的任意节点 (x, y) ,搜寻出所有其可能的前续节点,并选择其中累积距离最小节点作为前续节点,则到达该点路径的最小累积距离计算公式为:

$$D(x, y) = d(T(x), R(y)) + \min \{ D(x-1, y), D(x-1, y-1), D(x-1, y-2) \} \quad (3)$$

式中节点满足图 2 的局部路径限制方式,且测试语音特征矢量作为横轴,模板矢量作为纵轴。这样从

起点(1,1)开始搜索路径,利用式(2)反复递推后续节点的最小累积距离,经过 N 步后到止点(N,M)结束,即两矢量全长度最优路径匹配,则 $D(N,M)$ 为测试语音特征矢量与模板矢量的最小累积距离。

传统孤立词识别系统基于上述DTW算法,分别搜索出测试语音特征矢量与所有库模板矢量的最优匹配路径并计算出两者的最小累积距离 D ,然后选择其中 D 最小的模板所表示的语音作为判别结果^[8-9],具体算法流程如图3所示。

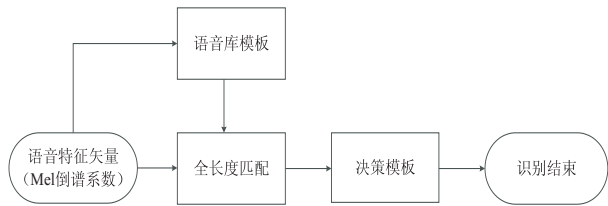


图3 传统全长度匹配流程

2 DTW 算法的改进

由上一节知道,传统孤立词语音识别时,需要将测试语音特征矢量与所有的库模板都进行全长度最优路径匹配。如果孤立词语音库模板量较大,系统的计算量将急剧上升,严重影响系统的识别效率。这个现象在一词多模板的非特定人识别系统中尤其突出。

2.1 改进思路

假设测试语音特征矢量表示的是实际孤立词 V ,则与模板库中其他孤立词模板相比,它的起始部分长度与模板库中同样表示孤立词 V 的模板的起始部分长度之间匹配失真度相对较小。反之,如果将测试语音特征矢量 V 的起始部分长度与所有库模板矢量进行最优路径匹配,那么只需要保留部分匹配失真度相对较小的库模板矢量就可以将表示孤立词 V 的模板矢量保留下来。由于匹配过程中只需要测试语音特征矢量的起始部分长度进行匹配运算,便可以排除匹配失真度较大的部分模板,因此系统减少了部分计算量。

综合以上想法,提出了测试语音特征矢量与库模板矢量部分长度最优路径匹配的DTW算法,具体思路如下:先截取提取到的测试语音特征矢量的起始部分长度(百分比),并从起点(1,1)开始搜索它们与各个模板矢量的最优匹配路径,采用松弛端点检测^[10]的方法找到最优匹配路径的止点并求出各自的最小累积距离 D ,即找出各模板矢量与截取语音特征矢量匹配度最大的起始部分长度并求出相应的 D ,然后选择 D 相对较小的部分模板进行保留,排除 D 相对较大的模板。如此循环,采用这种方法对剩余的模板进行部分长度匹配和排除,直至剩余模板数量唯一,该算法流程如图4所示。

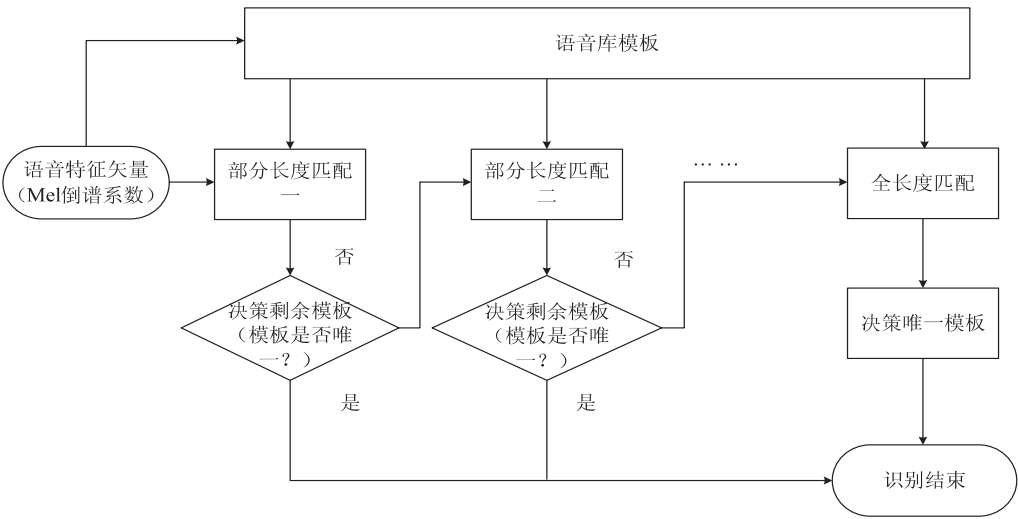


图4 部分长度匹配流程

2.2 准备工作

改进的DTW算法中测试语音特征矢量的截取方式和模板矢量保留方式需要预先设置,具体方法如下:每次截取的语音特征矢量都从起点开始,长度逐次增加。由于每次截取的测试语音特征矢量表示的语音可能与多个模板矢量起始部分表示的语音相同,所以每次部分长度匹配后保留的剩余模板一般大于1个(最后一次除外)。假设系统中有2个库模板矢量,分别表示语音“三”和“四”,它们的语音在起始端有相同的部

分:‘s’。如果截取的语音特征矢量表示的语音在‘s’范围内,一旦系统只保留一个模板,就有可能错判。为了保证部分长度匹配算法的识别率,每次部分长度匹配后采用如下方法估计保留模板个数:假设库模板总数为 l ,截取所有库模板矢量的起始部分,截取长度(百分比)与此次截取测试语音长度(百分比)一致,并将截取的各个模板矢量的起始部分表示为语音,则语音总数为 l 。选择其中一个截取语音,统计出库模板表示的语音中首部包含该语音的总数,记为该语音与

库模板语音的相似数 c_1 。同理,分别为其他截取语音统计出相似数,记为 $c_2 c_3 \cdots c_l$ 。比较所有相似数并取其中最大值 c_{\max} ,则 c_{\max} 与 l 的比值就是该次部分长度匹配后保留模板的百分比。

2.3 改进算法的步骤

改进 DTW 算法的步骤如下所述:

(1)将训练模板存入内存,总数记为 c ,同时进行识别阶段预设工作:设置测试语音特征矢量的截取方式,包括截取次数 m 和各次截取长度 a_1, a_2, \cdots, a_m (百分比);设置各次最优路径匹配后训练模板的保留个数 $b_1, b_2, \cdots, b_{m-1}$ (百分比且最后一次取一个模板, b_m 可忽略);

(2)输入测试语音信号,经过语音信号预处理(预加重、分帧加窗和端点检测)之后,提取出测试语音特征矢量;

(3)利用式(2)的条件排除部分训练模板,保留满足条件的训练模板;

(4)设某一保留训练模板矢量与测试语音特征矢量的帧匹配失真度矩阵为 d 和累积失真度矩阵为 $D = \text{Realmax}$,其中 d 和 D 的大小均为 $N * M$ 且横向表示待测语音帧,纵向表示训练模板帧。计算该训练模板矢量第一帧与测试语音特征矢量第一帧的失真度(欧氏距离),并保存到 $d(1,1)$ 和 $D(1,1)$ 中。同理,计算所有保留训练模板矢量第一帧与测试语音特征矢量第一帧的失真度(欧氏距离),并分别保存到各自的帧失真度矩阵与累积失真度矩阵相同的位置;

(5)搜索出训练模板矢量中与测试语音特征矢量第 $s \sim f$ 帧相交的且在图 1 平行四边形内的训练模板矢量帧,其中 $\begin{cases} s = a_{n-1} * N + 1 \\ f = a_n * N \end{cases}$ (四舍五入取整), n 为

截取迭代次数,初值 $n = 1$ 且 $a_0 = 0$;

(6)计算搜索到的训练模板矢量帧与待测语音 $s \sim f$ 帧之间的帧失真度(欧氏距离),并利用式(3)递推相交帧的累积失真度,分别保存到 d 与 D 相应的位置;

(7)找出累积失真度矩阵 f 列中最小的值,记为该训练模板矢量的最优部分匹配失真度 D_{\min} ;

(8)重复步骤 5~7,计算并找出所有保留训练模板矢量的最优部分匹配失真度 D_{\min} 。将保留的训练模板矢量按照各自的 D_{\min} 从小到大排序,并保留前 $c * b_n$ (四舍五入取整)个模板;

(9)判决 $c * b_n > 1$? 若是,则转入步骤 10 执行;若否,则转入步骤 11 执行;

(10)判决 $n < m$? 若是,则截取迭代次数 $n = n + 1$,转入步骤 5 执行;若否,则转入步骤 11 执行;

(11)将步骤 8 中已排序的训练模板矢量中的第

一个训练模板矢量表示的语音判决为测试语音,结束。

2.4 改进算法的识别性能评估

假设孤立词语音识别系统中有 N 个模板,平均测试语音提取参数的时间为 t_0 ,识别时间为 t_1 ,则全长度匹配需要的时间为 $N * (t_0 + t_1)$ 。在保证系统识别率基本不变的情况下,部分长度匹配需要的时间粗略计算如式(4):

$$T = N * [t_0 + t_1 * (a_1 + \sum_{i=1}^{m-1} (a_{i+1} - a_i) * b_i)] \quad (4)$$

其中, a_1, a_2, \cdots, a_m 逐次增加且 $b_1, b_2, \cdots, b_{m-1}$ 逐次减小。由式(4)可得, T 的值始终小于 $N * (t_0 + t_1)$ 。当 a_1 与 b_1 取值很小时,系统采用部分长度匹配算法所需的识别时间接近 $N * (t_0 + t_1 * (a_1 + b_1 - a_1 * b_1))$,远小于采用全长度匹配算法所需的时间 $N * (t_0 + t_1)$ 。

3 实验结果及分析

为了检验上述部分长度匹配 DTW 算法在降低系统识别时间方面的有效性,进行了两组仿真实验:比较部分长度匹配 DTW 算法与传统全长度匹配 DTW 算法,以及部分长度匹配 DTW 算法与文献[3]中增设参考模板阈值的 DTW 算法。实验中语音数据采自于 28 位男生,12 位女生,内容包括 0~9 十个汉语发音的数字。其中每个人对各个数字重复发音 2 次,总共 800 组数据。由于实验数据来自于非特定人,实验采用聚类的方法训练模板^[11-13],即对每个数字所有人的第一遍发音(40 组数据)进行聚类训练出 3 个或 6 个模板,10 个数字共 30 个或 60 个模板。其中语音参数采用 36 维 Mel 系数(12 维 Mel 系数+12 维一阶差分参数+12 维二阶差分参数)^[14-17]。语音采样频率为 16 kHz,预加重系数 $\mu = 0.9375$ 。采用汉明窗进行分帧,帧长 256 个样点,帧移 80 个样点。

实验中部分长度匹配 DTW 算法采用将测试语音特征矢量依次截取 3 部分(10%,70%,100%)的方式进行匹配,同时每次匹配后保留部分模板(20%,10%)作为下次部分匹配的参考模板,截取过程中矢量长度或模板个数的小数部分四舍五入取整。仿真结果如表 1 和表 2 所示。

从表 1 中的数据可以看出,在保证系统识别率基本不变的情况下,采用部分长度匹配算法的识别时间大约是传统全长度匹配 DTW 算法的 1/3,大幅降低了孤立词语音识别系统的识别时间。

从表 2 可以看出,相较于采用增设参考模板阈值的 DTW 算法,采用部分长度匹配 DTW 算法的识别时间也有明显降低。而且采用部分长度匹配 DTW 算法

避免了增设参考模板阈值的 DTW 算法额外设立合理阈值的问题,减少了识别系统的额外工作量。

表 1 部分长度匹配算法与传统 DTW 算法的识别性能比较

算法	参考模板个数	测试语音个数	识别率 /%	平均识别时间/ms
传统全长度	30	400	92.5	91.7
匹配 DTW 算法	60	400	95.5	158.5
部分长度匹	30	400	92.25	36.8
配 DTW 算法	60	400	96	53.4

表 2 部分长度匹配算法与文献[3]中 DTW 算法的识别性能比较

算法	参考模板个数	测试语音个数	识别率 /%	平均识别时间/ms
增设参考模板阈值	30	400	89.25	55.3
比较的 DTW 算法	60	400	95.75	96.1
部分长度匹配	30	400	92.25	36.8
DTW 算法	60	400	96	53.4

4 结束语

为了提高孤立词语音识别系统的识别效率,提出了测试语音特征矢量与模板矢量部分长度最优路径匹配的算法。在算法中,每一次进行部分长度匹配时只需截取一段测试语音特征矢量与剩余的模板矢量进行匹配,便可以大量排除匹配失真度较大的模板。仿真结果表明,与传统 DTW 算法相比,该算法在保证系统识别精度基本不变的情况下,减少了系统的工作运算量,有效降低了系统的识别时间。

参考文献:

[1] 余 炜,周 娅,万代立,等. 基于改进 DTW 的彝语孤立词识别研究[J]. 昆明理工大学学报:自然科学版,2014,39(5):47-53.

[2] ZHANG Z,TAVENARD R,BAILLY A,et al. Dynamic time warping under limited warping path length[J]. Information Sciences,2017,393:91-107.

[3] 张宝峰. 基于 DSP 的语音识别算法研究与实现[D]. 兰州:兰州理工大学,2011.

[4] 陈泉坤. 基于 DSP5509A 的 DTW 语音识别系统设计与实现[D]. 成都:电子科技大学,2012.

[5] 吴佳龙,李 坤,刘 中. 孤立词语音识别算法研究与设计[J]. 电子科技,2015,28(2):22-25.

[6] 陈立万. 基于语音识别系统中 DTW 算法改进技术研究[J]. 微计算机信息,2006,22(2-2):267-269.

[7] 朱淑琴,赵 瑛. DTW 语音识别算法研究与分析[J]. 微计算机信息,2012,28(5):150-151.

[8] 廖振东. 基于 DTW 的孤立词语音识别系统研究[D]. 昆明:云南大学,2015.

[9] 苏 昊,王 民,李 宝. 一种改进的 DTW 语音识别系统[J]. 中国西部科技,2011,10(1):38-39.

[10] 文 翰,黄国顺. 语音识别中 DTW 算法改进研究[J]. 微计算机信息,2010,26(7-1):195-197.

[11] ABDULLA W H,CHOW D,SIN G. Cross-words reference template for DTW-based speech recognition systems[C]//Tencon conference on convergent technologies for the Asia-pacific Region. [s. l.]:[s. n.],2003:1576-1579.

[12] 封伶俐,王秀萍. 一种新的基于 LBG 和 DTW 的模板训练算法[J]. 计算机工程与应用,2005,41(26):85-88.

[13] 李 燕,陶定元,林 乐. 基于 DTW 模型补偿的伪装语音说话人识别研究[J]. 计算机技术与发展,2017,27(1):93-96.

[14] 甄 斌,吴玺宏,刘志敏,等. 语音识别和说话人识别中各倒谱分量的相对重要性[J]. 北京大学学报:自然科学版,2001,37(3):371-378.

[15] DHINGRA S D,NIJHAWAN G,PANDIT P. Isolated speech recognition using MFCC And DTW[J]. International Journal of Advanced Research in Electrical Electronics & Instrumentation Engineering,2013,2(8):4085-4092.

[16] LIMKARA M,RAOB R,SAGVEKARC V. Isolated digit recognition using MFCC And DTW[J]. International Journal of Advanced Electrical & Electronics Engineering,2012,1(1):59-64.

[17] MUDA L,BEGAM M,ELAMVAZUTHI I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques[J]. Journal of Computing,2010,2(3):138-143.