

一种基于双帧动态时间规整的语音识别新方法

张亚歌, 张太镒

(西安交通大学 电子与信息工程学院, 陕西 西安 710049)

摘 要: 提出了一种基于双帧动态时间规整(Double Frame Dynamic Time Wrapping)的识别方法, 通过缩小帧匹配距离矩阵(Frame Match Distance Matrix)的规模从而减少计算量, 提高识别速度. 实验表明, 基于双帧动态时间规整的识别方法在识别率有所提高的条件下, 缩短了识别时间, 有很好的实用价值.

关键词: 语音识别; 双帧动态时间规整; 帧匹配距离矩阵

中图分类号: TP391.42

文献标识码: A

文章编号: 1000-7180(2010)11-0017-03

A New Method of Speech Recognition Based on Double Frame Dynamic Time Wrapping

ZHANG Ya-ge, ZHANG Tai-yi

(School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: To improve the recognition speed, a method of speech recognition based on Double Frame Dynamic Time Wrapping is proposed, this method decrease the size of Frame Match Distance Matrix and then increase the recognition speed. Experimental results show that the DFDTW method decrease recognition time with a better recognition rate and has practical value.

Key words: speech recognition; double frame dynamic time wrapping; frame match distance matrix

1 引言

语音识别技术就是让机器通过识别和理解把语音信号转变为相应的文本或命令的技术^[1], 其基本思路是提取语音信号中能代表语音内容的特征向量, 训练过程通过特定算法建立模板库^[2], 识别过程按照一定的规则将提取出的特征与模板库中的参考模板相对比得出识别结果.

语音识别系统一般包含有预处理模块、特征提取模块、训练模块和识别模块^[3]. 语音信号的短时平稳性决定了需要对其进行分帧处理, 预处理模块实现读入语音信号并对信号进行防混叠滤波、预加重、分帧、加窗、端点检测等功能; 特征提取模块针对每一帧进行运算, 提取出每一帧的特征向量, 帧语音信号提取出来的特征将是一个向量序列 $\mathbf{R} = \{\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_i, \dots, \mathbf{R}_T\}$, 其中分量 $\mathbf{R}_i$ 代表第 i 帧的特征

向量, 训练和识别过程均对特征向量序列(以下简称模板)进行操作; 训练模块对每条待识别词语的多个模板进行特定运算得到该待识别词语的模板后存入模板库; 识别模块将输入语音进行特征提取后与模板库中各个模板进行相似度匹配, 相似度最高的模板对应的词语作为识别结果. 典型的语音识别系统如图 1 所示.

语音识别系统的识别过程通常采用动态时间规整(DTW)、隐马尔可夫模型(HMM)、人工神经网络(ANN)等方法^[4], 尽管隐马尔可夫模型近年来已经成为识别方法的主流, 动态时间规整(DTW)以其简洁高效的特点仍在小词汇量系统中被广泛使用. 将 DTW 与其他方法结合进行识别的研究方兴未艾.

关于语音识别系统的性能, 对小词汇量孤立词识别系统来讲, DTW 识别速度要优于其他识别方法. 而 HMM 则具有高识别率的优势.

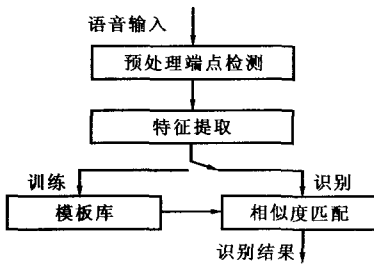


图1 典型语音识别系统模型

2 传统动态时间规整算法

动态时间规整 (Dynamic Time Wrapping, DTW)是语音识别中的经典算法,它利用动态规划 (dynamic programming)算法很好地解决了同一词语两次发音之间各音素速度不对应的问题.其基本原理如下所述:

设同一词语两次发音得到两个模板 $\mathbb{R}$ 和 $\mathbb{T}$, 分别为 M 帧和 N 帧向量序列,即

$$\mathbb{R} = \{R_1, R_2, \dots, R_m \dots, R_M\}$$

$$\mathbb{T} = \{T_1, T_2, \dots, T_n \dots, T_N\}$$

式中, R_m 为 $\mathbb{R}$ 第 m 帧的特征向量, T_n 为 $\mathbb{T}$ 第 n 帧的特征向量. 一般情况下 $M \neq N$, 且两次发音时各个音素单元并不是简单的线性对应关系, 因此, 要想求得两个模板之间的累积距离 (相似度匹配), 首先应该寻找一个规整函数 $i_m = \Phi(i_n)$, 将 $\mathbb{T}$ 的时间轴 $1, 2, \dots, i_n, \dots, N$ 非线性地映射到 $\mathbb{R}$ 的时间轴 $1, 2, \dots, i_n, \dots, M$ 上, 使得两个模板具有可匹配性. 图2描述了一个简单的规整函数.

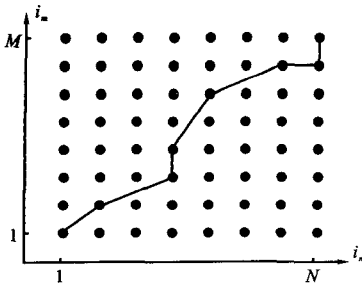


图2 规整函数示意图

一般说来, 规整函数应该满足:

$$D = \min_{\Phi(i_n)} \sum_{i_n=1}^N d(T(i_n), R(\Phi(i_n))) \quad (1)$$

式中, D 是在最优规整函数条件下两个向量序列的累积距离, $d(T(i_n), R(\Phi(i_n)))$ 是两个矢量 $T(i_n)$ 和 $R(\Phi(i_n))$ 之间的距离.

为了使匹配更准确, 规整函数需要有边界限制, 即

$$1 = \Phi(1), M = \Phi(N) \quad (2)$$

语音信号在时间上的顺序性又决定了规整函数必须满足单调性, 即

$$\Phi(i_{n1}) \geq \Phi(i_{n2}), \forall i_{n1} \geq i_{n2} \quad (3)$$

为了保证信息损失最小, 规整函数需要满足连续性, 即

$$|\Phi(i_n) - \Phi(i_n + 1)| \leq 1 \quad (4)$$

为了实现这些对规整函数的约束条件, 需要设计规整函数局部路径的约束, 典型的局部路径约束如图3所示.

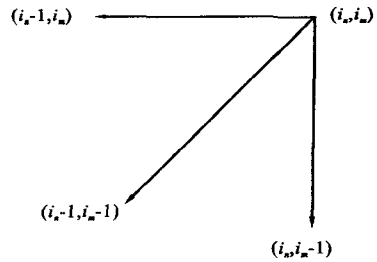


图3 典型局部约束路径

DTW 算法应用时步骤如下:

① 计算两个模板 $\mathbb{R}$ 和 $\mathbb{T}$ 各帧之间的距离, 得到帧匹配距离矩阵 $d(N, M)$. 矩阵中元素 $d(n, m)$ 按下式计算:

$$d(n, m) = d(T(n), R(m)) = \sqrt{\sum_{k=1}^K (T(n)_k - R(m)_k)^2} \quad (5)$$

式中, $T(n)_k$ 代表 $T(n)$ 的第 k 个分量, K 为特征向量的维数.

② 在帧匹配距离矩阵中按照局部约束路径找到一条最佳路径. 寻找最佳路径步骤为: 从匹配矩阵 $(1, 1)$ 点开始, 按照局部约束路径, 点 (i_n, i_m) 可达到的前一个节点为 $(i_n - 1, i_m)$ $(i_n - 1, i_m - 1)$ $(i_n, i_m - 1)$ 三者之一, 选择三者对应矩阵中最小值, 以此作为前一个节点. 此时路径累积距离为

$$D(i_n, i_m) = d(T(i_n), R(i_m)) + \min\{D((i_n - 1, i_m), D(i_n - 1, i_m - 1), D(i_n, i_m - 1))\} \quad (6)$$

递推至 (N, M) , 整个过程经历过的点的集合即为最优路径. $D(N, M)$ 即为最优路径的累积距离也即 $\mathbb{R}$ 和 $\mathbb{T}$ 的相似度.

③ 将待识别词的模板 $\mathbb{T}$ 与模板库中的 K 个参考模板 $\mathbb{R}^{(i)} (i = 1, 2, \dots, K)$ 进行逐一 DTW 匹配, 相似度最大的参考模板 $\mathbb{R}^{(i)}$ 对应的词条即判定为待识别词.

3 双帧动态时间规整

为了减少计算量,提高计算速度,本文提出了一种双帧动态时间规整法,基本算法流程如下所示。

① 将模板 $\mathbb{R}$ 和 $\mathbb{T}$ 每两帧的向量分为一组,得到双帧模板 $\mathbb{R}'$ 和 $\mathbb{T}'$:

$$\mathbb{R}' = \{\mathbf{R}'_1, \mathbf{R}'_2, \dots, \mathbf{R}'_m, \dots, \mathbf{R}'_M\}$$
$$\mathbb{T}' = \{\mathbf{T}'_1, \mathbf{T}'_2, \dots, \mathbf{T}'_n, \dots, \mathbf{T}'_N\}$$

式中, $\mathbf{R}'_1 = \{\mathbf{R}_1, \mathbf{R}_2\}$, $\mathbf{R}'_2 = \{\mathbf{R}_3, \mathbf{R}_4\}$, 依此类推, 如果 M 或 N 为奇数, 则在 $\mathbb{R}$ 或 $\mathbb{T}$ 尾部增加一个 $\mathbf{R}_M$ 或 $\mathbf{T}_N$ 再进行分组. 双帧模板 $\mathbb{R}'$ 和 $\mathbb{T}'$ 的组数 N' 和 M' 满足

$$N' = \text{ceil}(N/2), M' = \text{ceil}(M/2), \tag{7}$$

式中, $\text{ceil}(N/2)$ 表示不小于 $N/2$ 的最小整数。

② 计算双帧模板 $\mathbb{R}'$ 和 $\mathbb{T}'$ 的帧匹配距离矩阵 $d'(N', M')$ 。

$\mathbb{R}'$ 中双帧 $\mathbf{R}'_m$ 与 $\mathbb{T}'$ 中双帧 $\mathbf{T}'_n$ 的距离为 $d'(n, m)$ 定义为 $\mathbf{R}'_m$ 中两个向量分别与 $\mathbf{T}'_n$ 中两个向量的距离中的最小值. 即

$$d'(n, m) = \min\{d(T_{2n-1}, R_{2m-1}), d(T_{2n-1}, R_{2m}), d(T_{2n}, R_{2m-1}), d(T_{2n}, R_{2m})\} \tag{8}$$

式中, $d(T_{2n-1}, R_{2m-1})$ 表示原模板 $\mathbb{R}$ 中第 $2m-1$ 帧和 $\mathbb{T}$ 中第 $2n-1$ 帧之间的帧匹配距离, 按照式(5)计算。

③ 按照传统 DTW 算法在帧匹配距离矩阵 $d'(N', M')$ 中按照局部约束找到最佳路径. 最终得到识别结果。

由于 N' 和 M' 约是 N 和 M 的 $1/2$, 因此得到的帧匹配距离矩阵大小约为传统帧匹配距离矩阵的 $1/4$. 从而达到了减少计算量, 提高识别速度的目的。

4 实验结果

本实验在实验室微噪条件下进行对中文阿拉伯数字 1~10 进行训练和识别, 属于小词汇量孤立词语音识别, 实验信噪比约为 20dB, 每个数字的训练语音来自两男三女共 40 次发音, 帧长取 20ms, 帧间重叠 10ms, 特征向量使用语音信号的 12 维 mfcc 参数和 mfcc 参数一阶差分合并而成的 24 维向量, 训练过程采用 DTW 路径平均法得到模板库, 识别过程分别用 DTW、DFDTW、HMM 对两男三女共 1 000 组发音数据进行识别, 整个实验在 Matlab 平台上进行, 实验结果见表 1. 由实验结果可以看出, DFDTW 法识别 1 000 组词平均时间为 0.494s, 比传统 DTW 法的平均时间 0.813s 减少了 39.24%, 比

HMM 法的平均时间 0.902s 减少了 45.23%, 识别率方面比传统 DTW 提高了 3.2%, 比 HMM 低 1.1%. 整体看来, 新提出的 DFDTW 法在识别速度上有很大优势的同时对识别率的影响不大。

表 1 DTW、HMM 和 DFDTW 识别性能比较

识别词	DTW		HMM		DFDTW	
	识别时间 / 识别率		识别时间 / 识别率		识别时间 / 识别率	
	s	%	s	%	s	%
1	1.077	86.0	1.174	86.0	0.667	89.0
2	0.708	89.0	0.770	89.0	0.433	88.0
3	0.791	69.0	0.854	84.0	0.483	79.0
4	0.688	84.0	0.748	96.0	0.429	88.0
5	1.235	56.0	1.318	87.0	0.753	70.0
6	0.738	90.0	0.883	89.0	0.472	90.0
7	0.599	100.0	0.640	87.0	0.368	100.0
8	0.621	94.0	0.660	83.0	0.381	95.0
9	1.033	92.0	1.201	90.0	0.562	89.0
10	0.644	76.0	0.774	88.0	0.402	80.0
平均值	0.813	83.6	0.902	87.9	0.494	86.8

5 结束语

本文分析了语音识别领域中广泛运用的动态时间规整(DTW)法的原理, 提出了一种双帧动态时间规整(DFDTW)法, 实验表明, 在基本不影响识别率的情况下, DFDTW 比 DTW 和 HMM 有更快的识别速度, 节省了语音识别系统有限的计算资源. 本方法可结合语音增强^[5]、放松端点、减小搜索区域等改进的 DTW 法^[6]进行使用, 以得到更高的识别率和识别速度。

参考文献:

[1] Lawrence Rabiner, Biing-Hwang Juang. 语音识别基本原理[M]. 英文影印版. 北京:清华大学出版社, 1999.

[2] 冯晓亮, 于水源. 三种基于 DTW 的模板训练方法的比较[C]// 第八届全国人机语音通信学术会议论文集. 北京: 2005.

[3] Rabiner L R, Levinson S E. Isolated and connected word recognition - theory and practice[J]. IEEE Trans, 1981 (COM-29): 621-659.

[4] 吴金南, 宫宁生. 改进的 LVQ 网络与 DTW 相结合的语音识别方法[J]. 微电子学与计算机, 2009, 26(5): 33-35.

[5] 杨毅, 杨宇, 余达太. 语音增强及其消噪能力研究[J]. 微电子学与计算机, 2006, 23(7): 203-208.

[6] 付中华, 赵荣椿. 用窗口法在小存储器中实现 DTW 算法[J]. 西北工业大学学报, 2002, 20(4): 540-543.

生成得早,所以关键路径来自 C_{in} 和组间进位生成 g . 在布局布线中尽量应该将 C_{in} 和 g 信号放到与输出靠近的地方来减少路径延时. 很明显, nocarry 信号并不增加关键路径延时,因此采用这种办法设计最后的加法器.

6 结束语

本文设计的乘加器,采用了一种新颖的基-8 布斯编码,与基-4 的布斯编码相比,生成的部分积个数又有所减少,进一步压缩了部分积阵列. 由表 2 可知,基-8 布斯编码所生成的部分积较基-4 布斯编码少 $\frac{n}{6}$ 个数量级,当 n 越大时,基-8 布斯编码压缩部分积阵列的优势就越大. 所以该向量乘加器十分适合大位宽的多媒体向量运算.

表 2 编码

非布斯编码	基-4 布斯编码	基-8 布斯编码
n	$\frac{n}{6}$	$\frac{n}{3}$

在延时上,该基-8 布斯编码的向量乘加器在处理正负三倍被乘数的部分积时不需要进行传统的长进位链的进位传播加法,只需要一个 4 位的 CPA 延时,以及增加一项 EX2 部分积,较之传统的延时大大减小. 而且还引入进位消除(nocarry)信号,使得关键路径上减少了一个复用器,进一步降低了延时.

在面积上,基-8 布斯编码的向量乘加器比基-4 的更小,更适合苛刻条件下的高性能乘加运算,如航

天领域的处理器.

参考文献:

[1] 黄小平. 基于 Altivec 技术的向量处理单元的结构研究[D]. 西安:西北工业大学,2006:26-33.

[2] Neil H E Weste, David Harris. COMS 超大规模集成电路设计[D]. 3 版. 北京:中国电力出版社,2006:579-590.

[3] Dimitri Tan, Albert Danysh, Michael Liebelt. Multiple-precision fixed-point vector multiply-accumulator using shared segmentation[C]// IEEE Symposium on Computer Arithmetic. Spain, 2003.

[4] Shankar Krithivasan, Michael J Schulte. Multiplier architectures for media processing[C]// Proceeding of the 37th Asilomar Conference on Signals, Systems and Computer. Asilomar, USA. 2003:2193-2197.

[5] Richard P Brent, Paul Zimmermann. Modern research in computer arithmetic[M]. Cambridge University Press, 1997:34-57.

[6] 周旭. 低功耗浮点乘法部件的研究与设计[D]. 北京:中国科学院计算技术研究所,2005:17-42.

[7] 蒋小龙. 定点加法器设计[EB/OL]. [2002-09-14]. <http://www.pld.com.cn/ddvance/MuLD.Pdf>.

作者简介:

张喆 男,(1984-),硕士研究生. 研究方向为计算机系统结构、专用集成电路设计.

张盛兵 男,(1968-),博士生导师. 研究方向为 VLSI 系统设计、计算机系统结构、电子设计自动化.

黄小平 男,(1979-),博士研究生. 研究方向为计算机系统结构、专用集成电路设计.

(上接第 19 页)

作者简介:

张亚歌 男,(1984-),硕士研究生. 研究方向为语音信号处

理、语音识别.

张太镒 男,(1944-),教授,博士生导师. 研究方向为新一代移动通信技术、软件无线电、数字音视频技术.