一种新型的动态时间弯曲方法

A Novel DTW

钱小军　周建辉　潘雪增　平玲娣

（浙江大学计算机科学和工程学系　杭州310027）

Abstract　A key question in speech recognition is DTW. With the aims of solving the problem of dealing with the similar voices in classical DTW , this paper proposes a new DTW to solve the problem . The new DTW defines the global pattern dissimilarity as the maximum of the distances over all the possible paths. To make the new DTW meaningful, we must impose the constraints on it . On normal circumstances, the constraints include endpoint constraints, monotonicity constraints, local continuity constraints, weighting. At last, this paper points out the significance of the new DTW in the whole process of the speech recognition by showing the diagram of the process.

Keywords　DTW, Reference patterns, Test patterns, Feature extraction

1. 引语

在语音识别中有一个关键的问题就是语音模式如何比较以确定它们的相似性（或者称为等同性即模式之间的距离），特别在我们应用基于模板的语音识别方法的时候。在基于模板的系统中，如果从频谱对的差别或者说扭曲的角度来看，那么所使用的距离尺度有非常明显的物理意义。

语言可以直接由频谱向量的时间序列来表示，这些向量可以由前端频谱分析获得，这种分析模型有线性预测编码分析模型，过滤器组前端分析器模型。但是由于不同的人或者同一个人在不同的场合下发同一个音素时（包括单词，短语，句子），从整个发音过程来讲，其讲话的速度几乎是不可能完全相同的，所以我们需要将讲话的速度的抖动[1]进行标准化，以便在识别决定作出之前使发音的比较变得更加有意义。一般地，我们称这种标准化的方法为动态时间弯曲。

2. 经典的动态时间弯曲

如果我们使用 X 和 Y 来表示两个语音模式，其中 X 用时间序列(x_1, x_2, x_{T_x})来表示，Y 用时间序列(y_1, y_2, y_{T_y})来表示。我们用两个语言序列的下标指数，i_x 和 i_y，关联到一个共同的时间轴 k，那就是说：

$i_x=\Phi x(k), k=1,2,\cdots,T$

$i_y=\Phi y(k), k=1,2,\cdots,T.$

一个全局的模式差别尺度 $d_\Phi(X,Y)$可以定义为如下模式：

$$d_\Phi(X,Y)=\Sigma d(\Phi_x(k),\Phi_y(k))m(k)/M_\Phi \qquad (2.1)$$

在经典的动态时间弯曲中，我们将模式之间的距离 d(X,Y) 定义成 $d_\Phi(X,Y)$在所有可能路径上的最小值，也就是说：

$$d(X,Y)=\min d_\Phi(X,Y) \qquad (2.2)$$

事实上这个距离的定义有一些问题值得讨论。

直观来讲，当 X 和 Y 表达的是同一音素的发音时，上述定义是具有相当吸引力的，因为为了使沿着匹配路径累积的扭曲达到最小化而选择的最佳路径意味着这种扭曲是基于最优化的对齐策略而测量的距离。这种最优化的对齐策略其实补偿了在同一个音素的不同版本之间的非线性的说话速度的差别。但是当 X 和 Y 代表的是不同音素的发音时，方程(2.1)本身体现的对齐策略是最佳的这一点并不明显。因为我们不能保证当测试一个模式时，这个模式与它的参考模式之间的距离并不一定在所有的测试模式与参考模式之间的距离中是最小的。

我们可以想象，在一些特殊的情况下，上述思想会导致错误。特别是对于相似的音素。为解决这个问题，我们就考虑使用另外一种动态时间弯曲方法，我们称之为新型的动态时间弯曲方法。这种新型的动态时间弯曲方法的目标是计算为解决这 $d_\Phi(X,Y)$的最大值，这样就可以将不同音素的版本之间的距离进行扩大化，虽然同一音素的不同版本之间的距离同样会扩大。

但是如果我们选择适当的限制因素，我们可以确保后者的距离被控制在相对于前者较小的扩张程度。因为相似的音素经常在某些区域内变化巨大，而在另外一些地方变化较小，所以自然会使大的距离更大，也就是说，我们可以给距离乘以一个权数。

3. 新型的动态时间弯曲方法

首先我们指出，新型的动态时间弯曲模型跟经典的动态时间弯曲模型有许多相似性，所以我们可以说只要改变方程(2.1)和(2.2)就可以得到新型的动态时间弯曲。

在方程(2.1)中，我们可以将 m(k)的含义从斜率加权变为放大因子。如果说 $d(\Phi_x(k),\Phi_y(k))$是相对大的，那么这个因子将使这个距离变得更大；相反，如果说 $d(\Phi_x(k),\Phi_y(k))$是相对小的，那么这个因子的作用就是使这个距离变得更小。这样做的目的是使不同语音在变化较大的某些区域距离变得更大。

在方程(2.2)中，我们可以改变这个方程如下：

$$d(X,Y)=\max d_\Phi(X,Y) \qquad (3.1)$$

那就是说我们将得到的最大距离作为测试模式之间的距离。我们必须注意的是这种新型的动态时间弯曲是基于相似音素的。因此在识别阶段时，我们只要选择所有的 d(X,Y)中最小的距离，其相对应的参考模式就可以认定为测试模式所代表的音素。

钱小军　硕士，主要研究方向为智能信息处理与语音识别。潘雪增　教授，博导。平玲娣　教授，博导。

4. 新型动态时间弯曲的限制

为了使对于音素的不同版本的比较过程从时间标准化角度更为有意义，在这弯曲函数上加一些限制是非常必须的。如果对方程(3.1)中的弯曲函数不加任何限制的话，那么可以想象同一音素不同版本在这个最大化的过程中会导致不确定性，从而为了识别目的而做的比较变得毫无意义。

那些被认为对时间对齐有必要和合理的典型弯曲限制因素包括下面一些：

·端点限制

·单调性限制

·局部连续性限制

·加权

4.1 端点限制

当被比较的音素是要被识别的离散的音素时，它们通常由明确定义的端点来标识这个模式的开始帧和结束帧。如何决定或者说检测开始帧和结束帧在语音识别中是另外一个重要的问题，此问题被称为语言检测或者说端点检测。在本文中这个问题不是最关键的问题。当然如何做端点检测有许多方法，例如平均能量法。

因此我们可以讲一个模式的端点是事先给定的，时间的变化发生在由端点定义的范围里。因此，对于时间的标准化来讲，端点意味着固定的发音的时间限制，由此导致了如下形式的一套时间弯曲函数的限制：

开始帧　$\Phi x(1)=1$，　$\Phi y(1)=1$

结束帧　$\Phi x(T)=Tx$，　$\Phi y(T)=Ty$

4.2 单调性限制

正像以前讨论的一样，一个语言模式的频谱序列的时间顺序对于语言意义来讲是极其重要的。因此，在实施时间标准化时候，为了保持时间顺序，加上如下形式的单调性限制是非常合理的：

$\Phi x(k+1)>=\Phi x(k)$

$\Phi y(k+1)>=\Phi y(k)$

这种限制去除了沿着时间轴的倒转弯曲的可能性，即使在一个很小的时间间隔内。

4.3 局部连续性限制

在语言的发音中，有时一个特殊的声音或者说是音素的存在是一个唯一的可以识别的能方便正确地识别的因素。这一点对于相似音素的识别来讲是特别重要的，因此通过发现最佳的时间匹配的时间标准化方法不应该导致任何重要的载有信息的声音片段的丢失。

为了确保正确的时间对齐，同时使得任何潜在的信息丢失限制在一个最小的限度内，我们通常在时间弯曲函数上加入一系列的连续性限制。局部连续性限制有许多形式，其中的一个例子就是：

$\Phi x(k+1)-\Phi x(k)<=2$

$\Phi y(k+1)-\Phi y(k)<=2$

这样的约束规范通常是非常复杂的，由此如果用递增路径变化的方法来表达这些限制是非常方便的。但是递增路径变化的方法是另外一个重要的课题，具有极其丰富的内涵，这个不是本文的中心点。

4.4 加权系数

正如前边讨论的，加权函数 m(k)为寻找最优化的弯曲路径时增加了另外一个控制因素。

这条路径是为了使时间标准化反映语音模式中内在时间可变性而设置的，从而使得这条路径无论从语言上还是从听觉上都变得更加有意义。也就是说，加权函数 m(k)控制了每个短时距离 $d(\Phi x(k),\Phi y(k))$对整个距离和的贡献。

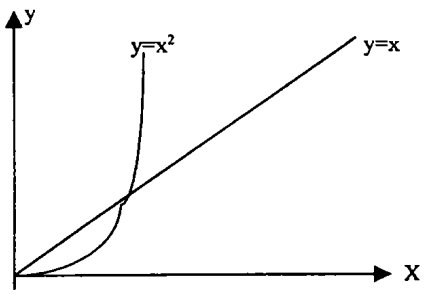

图1

首先我们分析一下上述的图形，我们可以看到如果$0=<x<=1$那么 $x*x<x$，反之如果 $x>=1$，那么 $x*x>x$，所以我们可以说如果我们将距离 $d(\Phi x(k),\Phi y(k))$进行平方，那么大于1的距离将会被扩大，小于1的距离将会被减少。当然，我们有许多种形式的 m(k)，那就是说我们可以设定一个阈值，高于这个阈值的距离将会被扩大，而低于这个阈值的距离将会被缩小。

但是又会存在着另外一个重要的因素，如果绝大多数的距离都高于一个特定的值或者绝大部分的距离都低于一个特定的值，那么我们必须对这个阈值做出必要的调整，否则的话我们不可能得到预想的结果。

事实上，我们可以首先计算所有的距离，然后选择所有距离的质心或者是中心作为阈值，接下来我们调整 m(k)，让它等于$(d(\Phi x(k),\Phi y(k)))^2$/阈值。

5. 新型的动态时间弯曲的实现

基于模板识别的语音识别的整个流程如下：

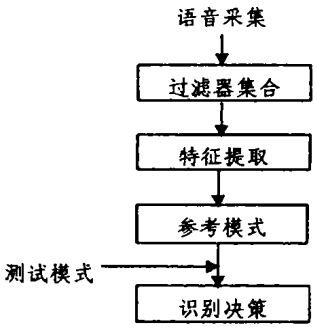

图2　系统流程图

对于上图，我们必须做出如下的说明：

1. 过滤器集合中，我们使用的是一个非同一的过滤器集合。有许多方式可以设计它，例如比较著名的关键频带方法，在低于1000赫兹的频率范围内它的频率幅度接近于线性的，而高于1000赫兹的频率是接近于对数的。但是我们使用的是关键频带的一个变种，即 mel 幅度方法；

2. 特征提取阶段中，我们使用的是快速傅立叶变换的方法来将时间域上的信号转变为频率域上的信号，同时我们使用的汉明窗口函数：$w(n)=0.54-0.46\cos(2*\Pi*n/(N-1))$；

（下转第62页）

```
//ga functions
public :
        void Cross(CSolution \&soluA, CSolution \&soluB);
        void Cross(int nNumber);
        void Mutation(int nNumber=5);
        void SortSolu();
        void CalcuErr(CSolution \&solu);
        double Get1.2Err(CSolution \&solu);
        double GetMaxErr(CSolution \&solu);
}
```

实验数据采用下列6种非线性项：

(1)超线性指数函数(SEN)。

$$f(t,x)=x^2\sin(t)\sin(x)+e^t\sin(x+3)+te^t+11$$

(2)次线性函数(SF)。

$$f(t,x)=x\sin(t)+5+\sqrt{|x|}$$

(3)超线性多项式(SP)。

$$f(t,x)=4x^4-2\sin(t)x^3+7x^2+xe^t-5$$

(4)弱奇性函数(WSN)。

$$f(t,x)=4x-\frac{2\sin(t)}{1-t}+xe^t+2$$

(5)强奇性函数(SSN)。

$$f(t,x)=tx-2\sin(t)+\frac{e^t}{x}-4$$

(6)一般非线性项(GSN)。

$$f(t,x)=t+2x^2+4\sin(tx)-\frac{\cos(t)}{x}+1$$

对上述非线性项，我们考虑下列问题：

$$\begin{cases}-x''(t)=f(t,x),t\in(0,1)\\ x(0)=x(1)=0\end{cases}$$

其中 population=20，变异参数为：

$$|x''(t)|\leqslant 10,\omega(x,t)\leqslant 0.1,|x''(0)|\leqslant 15,|x'(0)|\leqslant 10$$

最大误差为0.01。运行结果见下表。

表1 运行结果表

Nonlinear	Max Err	Last Err	Aver Err	Result
SEN	0.01	0.0039	14	OK
SF	0.01	0.0030	3	OK
SP	0.01	0.0099	15	OK
WSN	0.01	0.0083	25	OK
SSN	0.01	0.0078	14	OK
GSN	0.01	0.0020	423	OK

在本文最后，我们给出了系统初始图像，最终图像，错误曲线等。

致谢 本文是在作者访问香港理工大学期间完成的，资助项目有香港理工大学 Research Fellow Matching Fund Scheme 2001 (No. G. YY. 34)，国家自然科学基金，山东省中青年科学家奖励基金项目。

参 考 文 献

1 Ladde G S. Lakshmikantham V. Vatsala A S. Monotone iterative techniques for nonlinear differential equations. Pitman, 1985
2 Heikkila S. Lakshmikantham V. Monotone iterative techniques for discontinuous nonlinear differential equations. Marcel Dekker, Inc., New York, 1994
3 Agarwal R P, Chow Y M. Iterative methods for a fourth order boundary value ptoblems. J. Comput. Appl. Math., 1984 (10): 203～217
4 Buckles B P, Petry F E. Genetic algorithms. Los Alamitos, Calif.: IEEE Computer Society Press, 1992
5 Liu Xiyu. Some existence and nonexistence principles for a class of singular boundary value problems. Nonl. Anal, 1996, 27(10): 1147～1164
6 Guo D, Sun J. Nonlinear Integral Equations. Shandong Science and Technology Press, Ji-Nan, 1987
7 Guo Dajun, Lakshmikantham V. Nonlinear problems in abstract cones. Academic Press, New York, 1988
8 Goldberg D E. Genetic algorithms in search optimization and machine learning. Addison-Wesley, Reading, MA, 1989

(上接第70页)

3. 参考模式阶段中，距离的测量是平方差错误。当存在一个空元时，那个最密集的元包将会被分裂。第一个码本可以通过 N=1而获得；

4. 我们比较测试模式和参考模式的时候，上述的新型的动态时间弯曲方法将会被使用。因为比较的结果直接决定了语音识别的正确性，所以新型的动态时间弯曲是非常重要的；

5. 我们对语音的采样是采用单声道，采样频率是22050赫兹，文件格式采用.wav，去掉了文件头后，将数据经过适当的转化后读入相应的语音库文件；

6. 在识别阶段，我们可以先利用已有的经典动态时间进行衰别，然后设定一个阈值，低于这个阈值的语音被认为是相似语音，再将其通过新型的动态时间弯曲进行识别。

未来的工作 我们已经做的工作是完成了各个阶段的编程实现，我们将要做的工作就是对相似语音进行采样，并且对其分别采用经典的动态时间弯曲和新型的动态时间弯曲，再对其结果进行分析比较，看看结果是不是新型的动态时间弯的性能优于经典的动态时间弯曲。并对这样的结果进行比较分析，得出结论。虽然从理论上来讲后者的性能是优于前者的性能的。

参 考 文 献

1 Kaisheng Y, et al. Residual Noise Compensation For Robust Speech Recognition In Nonstationary Noise, IEEE ICASSP'2000, U. S. A, June 2000
2 邵央，冯哲，李宗葛. HMM 算法框架在银行语音服务中的实现. 计算机工程，2000，26(11)：126～129
3 Rabiner L, Juang B-H. 语音识别基本原理. 清华大学出版社，1999
4 郭军，等. 智能信息处理技术. 北京邮电大学出版社，1999