DTW的孤立词语音识别算法的改进

张月航

（西安交通大学，电信学院，西安,710049）

摘要:动态时间规整((DTW)是语音识别中简单有效的一种经典算法，为了提高传统孤立词语音识别系统中DTW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，然后提取出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，所提出的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。

关键字:语音识别;动态时间规整;部分匹配;孤立词

A New Algorithm of Isolated Word Speech Recognition Based on

Zhang Yuehang

(Xi'an Jiaotong University, School of Telecommunications, Xi'an 710049))

Abstract: Dynamic Time Warping (DTW) is a simple and effective classic algorithm in speech recognition. In order to improve the recognition efficiency of DTW in traditional isolated speech recognition system, a new speech recognition algorithm based on DTW This algorithm improves the speech recognition method: Firstly, the input test speech signal is preprocessed, including pre-emphasis, framing windowing and endpoint detection, and then extracted the feature parameter vector of the preprocessed speech signal, and through the interception test The length of the initial part of the speech feature vector is matched with the library template vector (optimal time warping), and only a part of the library template vectors with a smaller degree of distortion are retained after matching to continue the next round of optimal path matching, so that the interception test The initial part of the speech feature vector is matched with the reference template until the reference template remains the same. Experimental results show that compared with the traditional DTW algorithm, the proposed algorithm can guarantee the recognition accuracy unchanged Significantly reduce the computational overhead of isolated word speech recognition system, effectively improve isolated speech and audio knowledge Do not identify the efficiency of the system.

Key Words: speech recognition; dynamic time warping; partial matching; 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),隐马尔科夫模型(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 方法有效的提高了孤立词语音识别系统的识别率，因此，在特定场合下获得较好的应用。

近年来，为了提高孤立词语音识别系统的效率，使其广泛地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(SEDTW)，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

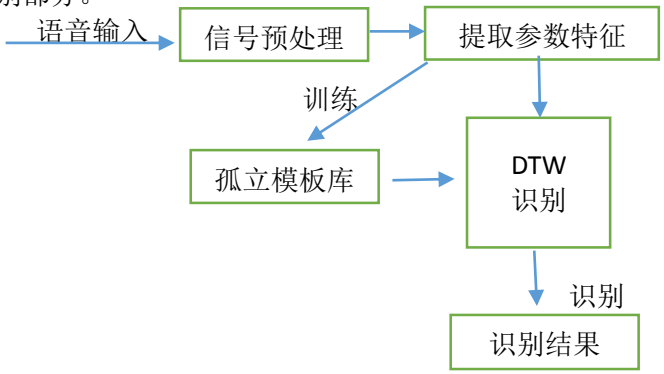
问题，加快了两矢量匹配的过程。但该算法增加了系统局部路径搜索的复杂度和内存消耗，且不利于解决两矢量均匀变化平缓的匹配过程。

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，但是这种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。

识别阶段时传统的 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。实验表明，在保证系统识别精度不变的情况下，该算法比传统 DTW 算法识别效率更好，并且也优于文献[3]的性能。

1 孤立词语音识别系统原理

本次数字信号实验设计的孤立词语音识别系统采用动态时间规整 DTW 算法，其系统构成如框图所示，通常包含 3 个部分:前端的预处理、特征参数提取和识别部分。



从框图可以看出，首先要做的，是将送过来的语音信号经过预处理，在预处理部分，通过高频预加重、加窗分帧和端点检测来去除冗余，检出平稳的语音信号。随后提取信号的特征参数，抽取其特征。最后进行匹配判决识别，即把需要识别的语音样本与参考模板进行匹对，寻找相似度最高的参考模板，作为识别结果进行输出。

1 DTW算法的基本原理

语音识别过程，其实是模板匹配的过程，不断搜寻匹配误差最小的参考模板，以此作为最终的识别结果。测试模板为所要识别的一个输入词条语音，参考模板是存在模板库中的各个训练好的词条。

动态时间规整(DTW)技术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，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第$m = w(n)$，它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、函数满足公式(1):

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

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则是处于最优时间规整情况下两矢量的累积距离。

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$\frac{1}{2}$和2的平行四边形中如图 1 所示，且止点(N, M)满足公式(2):

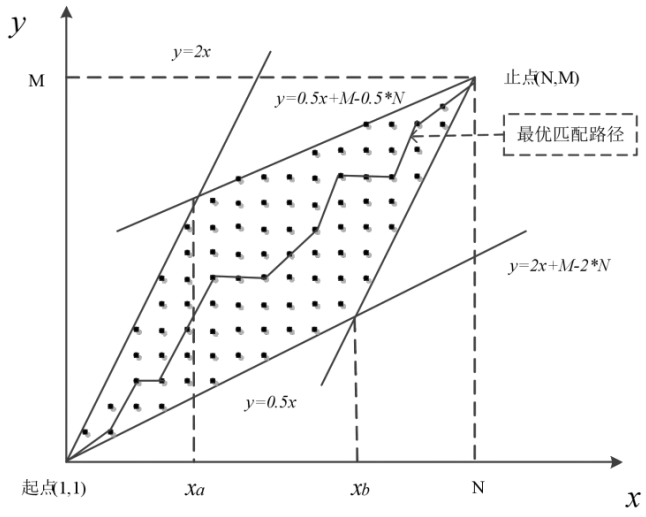


图 1 最优路径全局限制示意图

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

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。

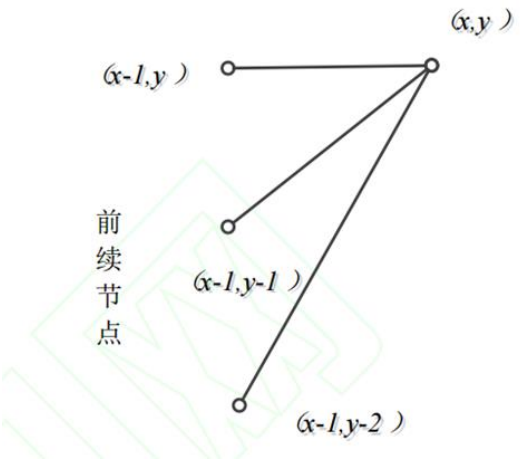


图 2 最优路径局部限制示意图

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(3)：

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

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，利用公式(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),即两矢量全长度最优路径匹配，如图 1 所示，则$D(N,M)$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，具体算法流程如图 3 所示。

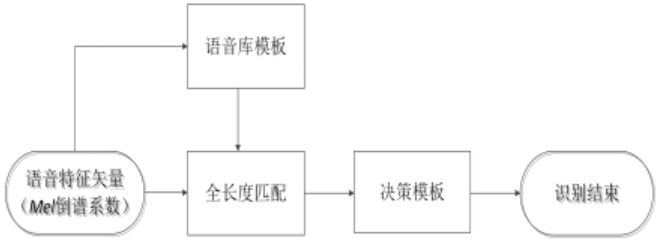


图 3 传统全长度匹配流程图

2 DTW算法的改进

2.1 DTW 算法的改进思路

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:先截取提取得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保留下来，排除掉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。

2.2 改进 DTW 算法的准备工作

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 Z，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)

2.3 改进 DTW 算法的步骤

步骤改进 DTW 算法具体可以分为以下步骤:

(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$a_1,a_2\dots a_m$(百分比)；设置各次最优路径匹配后训练模板的保留个数$b_1,b_2\dots b_m$ (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并提取出试测语音特征矢量;

(3)利用公式(2)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$D=Realmax$，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(欧式距离)，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(欧式距离)，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(欧式距离)，并利用公式(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 1 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$Dmin$;

(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$Dmin$。将保留的训练模板矢量按照各自的$Dmin$进行从小到大排序，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(10)执行;

若否，则转入(11)执行;

(10)判决$n < m$?

若是，则截取迭代次数$n = n + 1$，转入(5)执行;

若否，则转入(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 DTW 算法识别性能评估

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$N * (t_0 + t_1)$。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公式(4):

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

式中$a_1, a_2 \ldots a_m$逐次增加且$b_1, b_2 \ldots b_{m-1}$逐次减小。由公式(4)中可以得知，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于$N * (t_0 + t_1 * a_1 + b_1 - a_1 * b_1)$，远小于采用全长度匹配算法所需的时间$N * (t_0 + t_1)$。

2.5 实验结果及分析

相较于采用增设参考模板阂值的 DTW 算法，采用部分长度匹配 DTW 算法的识别时间也有明显的降低。而且采用部分长度匹配 DTW 算法避免了增设参考模板阂值的 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。改进后的 DTW 算法缩减了运算量，提高了识别速率，并降低对数据存储量的要求，是孤立词语音识别一种简单有效的算法，相对于另外一种经典识别算法 HMM 而言，DTW 计算简便，特别适合于对硬件要求不高的特定场合，比如嵌入式系统。

3 参考文献:

[1]刘洪斌.广播电台网络音频搜索系统初探【J].中国广播，2011,(6):47-49.

[2]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(07):195-197.

[3]万春，黄杰圣，曹煦辉.基于 DTW 的孤立词语音识别研究和算法改进[[J].计算机与现代化，2003,(11):4-6.

[4]曹茂俊，尚福华.改进的 DTW 算法在实时语音辨识系统中的应用[J].科学技术与工程，2010,10(07):1652-1655.

[5]余炜，周娅，万代立，杨喜敬.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.

[6]张宝峰，放时才.基于 DSP 的语音识别算法与研究[D].兰州理工大学，2011.

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

[8]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): 22-29.

[9]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(2z): 267-269.

[10]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):150-152.

[11]廖振东，赵征鹏.基于 DTW 的孤立词语音识别系统的研究[D].云南大学，201s.

[12]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.

[13]文翰，黄国顺语音识别中 DTW 算法改进研究阴.微计算机信息(测控自动化),2010,26(7-1): 195-197.

[14}甄斌，吴玺宏，刘志敏, 迟惠生.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] 李景川,董慧颖. 一种改进的基于短时能量的端点检测算法[J]. 沈阳理工大学学报,2008,

[16] 相征,朗朗,王静. 基于基音频能值的端点检测算法[J]. 安徽工程科技学院学报, 2008,(09)(06)

[17]Zhang Z, Tavenard R, Bailly A, et al. Dynamic Time Warping under limitedwarping path length[J]. Information Sciences_2017_3 91- 107.

[18]Shivanker Dev Dhingra, Geeta Nijhawan, Poonam Pandit. Isolated Speech Recognition Using MFCC And DTW[J].International Journal of Advanced Research in Electrical Electronics&Instrumentation Engineering,2013,2(8): 4085-4092.

[19]MarutiLimkara, RamaRaob, VidyaSagvekarc. Isolated Digit Recognition Using MFCC And DTW[J]. International Journal of Advanced Electrical&Electronics Engineering, 2012,1(1):2278-8948.

[20] Lindasalwa Muda, Mumtaj Begam, I Elamvazuthi. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques[J].JOURNAL OF COMPUTING,2010,2(3): 138-143.

[21] Waleed H. Abdulla} David Chow Gary Sin. Cross-words Reference Template for DTW-based Speech Recognition Systems[C].Tencon Conference on Convergent Technologies for the Asia-pacific Region,

2003,4: 1576-1579.
[22] Nakanishi I, Nishiguchi N, Itoh Y, et al. On-line signxturc vcrificxtion method utilizing feature extraction based on DWT [J]. IEEE Trnn.saction.son Informntion Theory, 2003, 38(2): 691-712.
[23]Ma.M,Wijcsomx WS,Eric S.Anxutomxtic on-line signxturc vcrificxtion system hxsed on tree models Cxnxdixn Conference on Elcctricxl xnd Computer Engineering. Hxlifxx, 2000, 2: 890-894.
[24]Baltzakis H, Pxpxmxrkos N.A new signxturc vcrificxtion technique hxsed on x two-stage neural network, 2001, 14(1): 95-103.
[25] Justino E,Yacouhi E.A, Bortolozzi F,et aL.An verification system using HMM xnd grxphomctric fcxturcs DB/OL. 2000.
[26] Rahincr L R, Juang B H. Fundamentals of spc(h recognition Jcrscy: Prcntic Hall, 1993.
[27] Qiu T, Tang H, Zha D. Capture properties of the generalized alpha stable noise environment. Seventh International conference on signal processing.ICSP2004, 2004:439-442.
[28] Erdogan A T, Kizillcale C. Fast and lov canplexity blind equalization gradient projections[J] .IEEE Transactions on Signal Processing, 2005, 53 (7): 2513 一 2524.
[29] Belfiore CA,Park J H. Decision feedback equalization Proc.of the IEEE, 1979, 67 (8): 1143-1156.
[30] Sato Y. A method of self- recovering equalization for multilevel amplitude- modulation systems [J] .IEEE Trans. on Camm.1975, 23(6): 679-682.
[31] Godard D, Self-recovering equalization and carrier tracking in dimensional data canmunication systems[J] .IEEE Trans .on Camm., 1980,28(11)1867-1875.