

文章编号: 1001-9081(2013)11-3138-03

doi: 10.11772/j.issn.1001-9081.2013.11.3138

基于模板匹配的快速语音关键词检出方法

朱国腾¹, 孙伟^{2*}

(1. 中山大学 信息科学与技术学院, 广州 510006; 2. 中山大学 软件学院, 广州 510006)

(*通信作者电子邮箱 sunwei@mail.sysu.edu.cn)

摘要: 在缺乏训练样本的情况下对语音信号进行关键词检出, 基于模板匹配的方法与传统的方法相比, 仍然能够对语音进行关键词检出。但是由于模板匹配方法计算局部最小距离的方式是逐帧移动, 所以计算时间长。局部最小距离的极值点通常在音素分割点附近, 利用这两者的位置关系并结合插值思想, 提出一种快速的模板匹配方法。该方法通过在音素分割点之间插值计算局部最小距离, 能够有效地缩短计算时间。在 TIMIT 和 CASIA 语料库中进行实验, 改进的方法与常规的模板匹配方法相比较, 快了约 2.8 倍。

关键词: 关键词检出; 动态时间规整; 音素分割; 插值

中图分类号: TN912.34; TP391.42 **文献标志码:** A

Rapid speech keyword spotting method based on template matching

ZHU Guoteng¹, SUN Wei^{2*}

(1. School of Information Science and Technology, Sun Yat-sen University, Guangzhou Guangdong 510006, China;

2. School of Software, Sun Yat-sen University, Guangzhou Guangdong 510006, China)

Abstract: When dealing with keywords detection without training samples, template matching-based keyword spotting can still be able to spot compared with the traditional method. However, template matching-based method is time-consuming, because it uses frame-by-frame move method to calculate the local minimum distance. The extreme points of the local minimum distance are usually near phoneme segmentation points. A fast template matching method can come out by combining their positions with interpolation idea. By using interpolation to generate the local minimum distance between phoneme segmentation points, this method can greatly reduce the calculation time. When running on the TIMIT and CASIA corpus, the improved method approximately is 2.8 times faster than the conventional template matching-based keyword spotting.

Key words: keyword spotting; Dynamic Time Warping (DTW); phoneme segmentation; interpolation

0 引言

语音关键词检出 (Keyword Spotting, KWS) 是在一段连续语音中检测是否出现了对应的关键词的发音的过程。关键词检出一直是语音识别研究的热点, 在安防监听、语音通信、话题跟踪等领域中都有重要的应用。

目前的语音关键词检出方法主要有两种^[1]: 一种是将所有语音都识别成文字, 然后对文字进行关键词检出; 另一种是基于垃圾模型或基于 Lattice 的关键词检出系统。这两种方法的声学模型复杂, 并且训练过程需要使用具有详细标注的发音样本。以上方法在缺少训练语音样本数据的情况下, 如方言、最新流行词汇, 难以进行快速精确的关键词检出。语音关键词检出通常是采用隐马尔可夫模型^[2] (Hidden Mark Model, HMM) 来实现, 而基于 HMM 的关键词检出存在以下问题^[3]: 一方面, HMM 模型需要大量发音样本进行训练, 更精确的识别需要样本细分到音节或音素级别; 另一方面, 模型的稳定性弱, 新加入的关键词需要重新进行训练。

针对传统方法的不足, 文献[4]提出了一种基于模板匹配的关键词检出方法, 该方法不需要任何训练数据和复杂的声学模型, 只需要一个发音样本作为模板, 通过动态时间规整 (Dynamic Time Warping, DTW)^[5] 计算局部最小距离 (Local Minimum Distance, LMD), 并自动确定检出阈值。基于模板匹

配的关键词检出方法在缺乏训练样本的情况下, 能取得比垃圾模型更高的召回率^[4]。但是, 该方法在每一帧中都要进行窗口内部滑动, 每一次滑动都使用 DTW 计算局部最小距离, 而 DTW 是一种动态规划算法^[5], 动态规划的时间复杂度和空间复杂度都相对较高, 因此, 模板匹配方法所需的计算时间太长, 故而不适于很多实时应用。

本文提出了一种快速的模板匹配方法, 即采用插值方法得出音素分割点内的 LMD, 从而减少计算时间。在分析了局部最小距离与音素分割点的位置关系后, 提出了查找潜在音素分割点的步骤, 最后通过在查找到的音素分割点之间插值得出全部 LMD 值。实验结果显示本文提出的方法所需要的计算时间少于常规的模板匹配方法^[4,6] 的方法。

1 基于模板匹配的关键词检出

模板匹配方法的模板可以是关键词的任意一次发音, 对待检测语音的每一帧都使用 DTW 去计算模板同滑动窗之间的 LMD, 利用局部最小距离出现次数的统计特性来确定阈值。

基于模板匹配的关键词检出算法流程^[4] 如图 1 所示。首先对关键词的模板语音段和待检测的语音段提取特征。然后, 在待检测语音段中逐帧移动提取滑动窗, 利用 DTW 计算关键词特征和滑动窗特征之间的 LMD。最后, 用 LMD 各个

收稿日期: 2013-04-24; 修回日期: 2013-07-02。

作者简介: 朱国腾(1988-), 男, 湖南郴州人, 硕士研究生, 主要研究方向: 模式识别、语音处理; 孙伟(1972-), 男, 江苏连云港人, 教授, 博士生导师, 主要研究方向: 多媒体安全、数字媒体。

距离的出现次数的统计特性估计检出阈值,如果待检测语音段中存在连续几帧的 LMD 值都低于检出阈值,就判断待检测语音包含关键词。

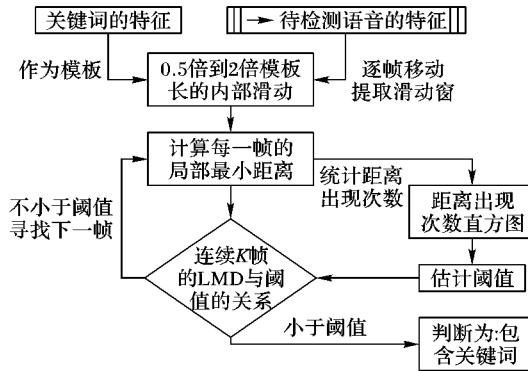


图 1 模板匹配关键词检出算法流程

对同一个词的每次发音,发音长度的范围不固定。窗口(Window, W)的大小在 0.5 倍模板长(Mold, M)和 2 倍模板长之间,才能够包含该关键词(Template, T)的发音范围^[4]。每一次窗口滑动都用 DTW 方法计算窗口同关键词特征之间的距离,从而确保找到局部最小距离。对第 n 帧的局部最小距离 LMD(n) 采用如下公式计算:

$$LMD(n) = \min_{0.5M < i < 2M} \{ DTW(T, W(i)) \} \quad (1)$$

其中:M 为模板长度,i 为窗口的滑动大小,滑动范围在 0.5 倍模板长和 2 倍模板长之间。图 2 是每一帧与对应局部最小距离的示意图。

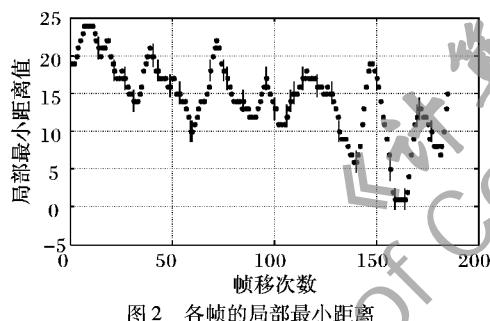


图 2 各帧的局部最小距离

对局部最小距离的各个距离值出现次数进行统计,得到距离出现次数直方图。检出阈值可以采用局部最小距离出现次数直方图的众数和标准差来估计^[4]。如果局部最小距离中存在连续的 k 帧^[6]都小于阈值,就判断该发音句子包含关键词,这 k 帧即为关键词在语音段中出现位置。

2 快速的模板匹配方法

由于同一音素内语音段发音存在连续性^[7],故滑动窗的局部最小距离的分布的波峰波谷或者关键转折点通常在语音分割处取得。图 2 中线段部分为音素分割点所在位置,音素分割点通常分布在局部最小距离的转折点或极值点处。

查找语音信号的音素分割点常用的特征参数有:音量幅值(Volume)、短时能量、短时过零率(Zero Cross Ratio, ZCR)和短时自相关等特征参数^[8]。其中音量幅值和短时过零率是区分清音和浊音^[9]最常用的两种特征。本文的音素查找结果仅需要尽可能多地覆盖音素分割点的大概位置,于是选择音量幅值和短时过零率的极值点,来查找语音信号的音素分割点^[7,9]。采用如下步骤:

1) 预处理。为方便计算 Volume 和 ZCR,首先将语音信号重采样为 8 kHz,通过端点检测提取发音段;然后进行零点调

整,并归一化幅度到 [-1,1] 的范围内;最后按照帧长 256,帧重叠为帧长的一半,进行分帧处理。

2) 提取参数。计算各帧的 Volume 和 ZCR,并分别提取所有帧的 Volume 和 ZCR 的极值点的位置。

3) 合并。把上一步提取到的两组极值的位置点合并为一组位置点。将这组位置点中距离太近的点合并为一个点,即对距离间隔小于一定阈值的一组点,用它们的中心点代替这组点。最后得到的位置点作为要找的音素分割点位置。

通过以上步骤,可得到音素分割点的大概位置,称这样的位置点为潜在音素分割点。

图 3 中横轴下部分线段所示为计算得到的潜在音素分割点位置,上半部分线段为实际音素分割点的位置,线段之间的字母是这部分发音的音素。对于潜在音素分割点密集出现的情况,可以在第 3) 步中缩小合并阈值,从而减少密集出现次数。

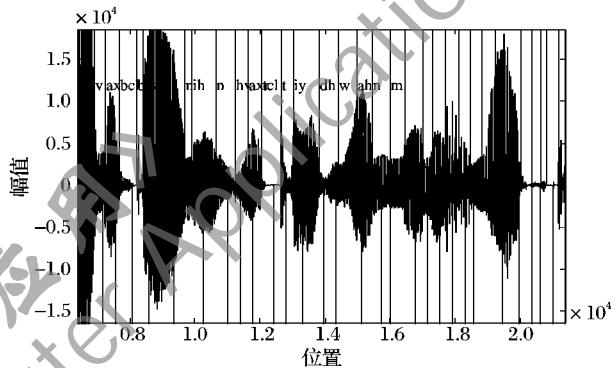


图 3 查找到的音素分割点和实际位置对比

对潜在音素分割点处使用模板匹配方法计算 LMD,而潜在音素分割点之间的 LMD 则通过插值得到。图 4 是逐帧移动和插值方法得到的 LMD 的对比,插值方法分别使用线性插值和三次样条插值。由于样条插值使用分段多项式进行插值^[10],多项式对插值点的波峰波谷或转折点有更好的逼近,故样条插值与线性插值相比,产生的 LMD 更接近逐帧移动计算的结果。

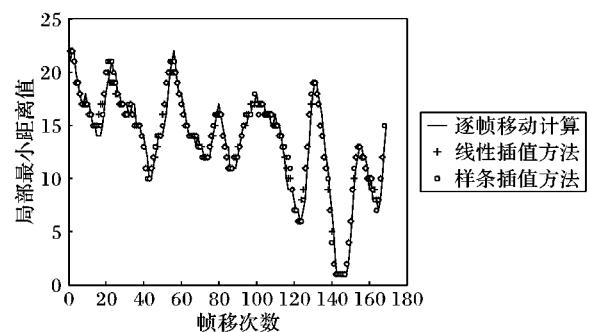


图 4 逐帧移动和插值方法的 LMD 对比

3 实验结果

实验所用语料库分别为:TIMIT 库和 CASIA 库。TIMIT 是来自美国 8 个不同方言地区的 6300 句发音,包含 6000 多个英文词汇^[11]。CASIA 库是由中国科学院录制的汉语语音合成语料库,包含 4000 句男播员的发音。

将实验所用到的语音信号重新采样到 8 kHz,进行端点检测提取发音段,使用 0.95 的因子进行预加重,使用 14 维的 Mel 频标倒谱系数(Mel Frequency Cepstrum Coefficients, MFCC)作为语音特征。实验环境为 Windows 7 系统、CPU 为 3.20 GHz、内存为 3.75 GB 的计算机,在 Matlab 软件中运行。

逐帧移动方法按照文献[4]的描述计算,插值方法分别选用样条插值和三次样条插值计算,阈值的参数 c 取 1,连续帧的参数 k 取 4,窗口在 0.5 倍模板长和 2 倍模板长之间内部滑动 10 次。

3.1 关键词检出性能的评价

关键词检出的主要目标是快速准确地鉴别待检测语音句子是否存在关键词的发音,检出情况有以下四种:TP,表示包含关键词发音的句子检测结果为包含;FP,表示不包含关键词发音的句子检测结果为包含;FN,表示包含关键词发音的句子检测结果为不包含;TN,表示不包含关键词发音的句子检测结果为不包含。

基于以上四种检出情况可以定义下面三种性能指标^[12]:

$$\text{查准率} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{召回率} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{正确率} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (4)$$

对每个关键词的查准率、召回率和精确度分别进行平均,得到平均查准率、平均召回率和平均正确率。再考虑每个句子检出所花费的时间,将每个句子的平均计算时间作为速度评价指标。采用 Matlab 的 etime 函数来计算时间(单位为 s)。

3.2 实验结果及分析

从 TIMIT 库中随机地选择 10 个单词发音作为关键词,分别为:word、voice、paper、home、mean、black、social、movies、greasy、water。对每个关键词从 TIMIT 库中随机地选择 5 句包含关键词发音的句子,及 50 句不包含关键词的句子进行实验。表 1 为在 TIMIT 库中逐帧移动、线性插值和三次样条插值三种方法的检出效果。

表 1 TIMIT 语音库实验结果统计

评价指标	文献[4,6]方法		本文方法
	逐帧移动	线性插值	样条插值
平均查准率	0.7286	0.6529	0.6957
平均召回率	0.5856	0.6350	0.6067
平均正确率	0.7880	0.7400	0.7680
平均计算时间/s	33.3448	11.8143	12.1498

同理,在 CASIA 库中随机选择 10 个词发音作为关键词,分别是:两国、问题、农村、朋友、百分、发展、社会、增加、合作、起了;并随机选择 5 句含关键词发音的句子,及 50 句不包含关键词的句子进行实验。表 2 为三种方法的检出效果。

表 2 CASIA 语音库实验结果统计

评价指标	文献[4,6]方法		本文方法
	逐帧移动	线性插值	样条插值
平均查准率	0.6314	0.5920	0.6029
平均召回率	0.7474	0.8333	0.8712
平均正确率	0.8619	0.8450	0.8510
平均计算时间/s	42.2653	14.9586	15.2225

实验结果表明插值方法极大地缩短了计算时间,比文献[4,6]中常规的逐帧移动方法快了 2.8 倍。样条插值比线性插值花费的时间稍微多一些,是因为样条插值需要使用多项式函数去计算每一段的多项式系数。

插值方法与逐帧计算相比查准率和正确率略低,而召回率较高。这是因为插值方法产生的局部最小距离出现次数的方差小,从而得到的检出阈值大,更多的发音句子被识别结果为包含关键词,也即与逐帧计算的结果相比 TP 和 FP 都变大,相应的 FN 和 TN 变小,而实验所用的负样本数量比正样本多,故而 FP 和 TN 的变化幅度比 TP 和 FN 的变化幅度大。

实际操作中,可以通过改变阈值的参数 c 来调整查准率和召回率的关系^[4]。

综合考虑四种评价指标,得出样条插值方法更适合基于模板匹配的快速语音关键词检出。

4 结语

在训练数据缺乏的情况下,传统算法难以对语音关键词进行检出,而模板匹配的方法不需要训练数据和复杂的声学模型,仅用一个模板的发音就可以进行关键词检出。但是模板匹配方法采用逐帧移动去计算模板和窗口的局部最小距离,检测过程需要大量的计算时间。本文针对模板匹配方法费时长的问题进行改进:首先分析了音素分割点和局部最小距离两者的位置关系;然后提出了查找音素分割点的方法;最后用插值方法得到全部 LMD。实验结果表明,插值方法与原方法相比,在保证查准率、召回率和准确率没有很大变化的同时,可以大幅缩短计算时间。总之,基于模板匹配的快速关键词检出方法,在训练数据缺乏的条件下,仍然能够快速地对关键词进行检出。未来的工作可以围绕快速 DTW 算法或寻找更精确的音素分割点等方面来进行。

参考文献:

- [1] 韩纪庆, 张磊, 郑铁然. 语音信号处理[M]. 北京: 清华大学出版社, 2004.
- [2] KESHET J, GRANGIER D, BENGIO S. Discriminative keyword spotting[EB/OL]. [2013-03-20]. <http://eprints.pascal-network.org/archive/00003299/02/KeshetGrBe07.pdf>.
- [3] ROSE R C, PAUL D B. A hidden Markov model based keyword recognition system [C]// ICASSP'90: Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing. Albuquerque: Albuquerque Convention Center, 1990: 129 - 132.
- [4] BARAKAT M S, RITZ C H, STIRLING D A. Keyword spotting based on the analysis of template matching distances [C]// ICSPCS 2011: Proceedings of the 5th International Conference on Signal Processing and Communication Systems. New York: IEEE Communications Society, 2011, 1 - 6.
- [5] SAKOE H, CHIBA S. Dynamic programming algorithm optimization for spoken word recognition [J] IEEE Transactions on Acoustics Speech and Signal Processing, 1978, 26(1): 43 - 49.
- [6] BARAKAT M S, RITZ C H, STIRLING D A. Detecting offensive user video blogs: an adaptive keyword spotting approach [C]// ICALIP 2012: Proceedings of the 2012 International Conference on Audio, Language and Image Processing. Washington, DC: IEEE Computer Society, 2012, 419 - 425.
- [7] 张江安, 杨洪柏, 林良明, 等. 一种基于段间距离测度的语音自动分割方法[J]. 上海交通大学学报, 2001, 35(9): 1362 - 1365.
- [8] 屈丹, 王波, 李弼程. VoIP 语音处理与识别[M]. 北京: 国防工业出版社, 2010.
- [9] 王宁, 万旺根, 余小清. 汉语语音音素分割的一种新方法[J]. 上海大学学报: 自然科学版, 2002, 8(2): 116 - 118.
- [10] 张诚坚, 高健, 何南忠. 计算方法[M]. 北京: 高等教育出版社, 1999.
- [11] GAROFOLO J S, LAMEL L F. TIMIT acoustic-phonetic continuous speech corpus, 2013[EB/OL]. [2013-03-21]. <http://www.ldc.upenn.edu/Catalog>.
- [12] POWERS D M W. Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation[J]. Journal of Machine Learning Technologies, 2011, 2(1): 37 - 63.