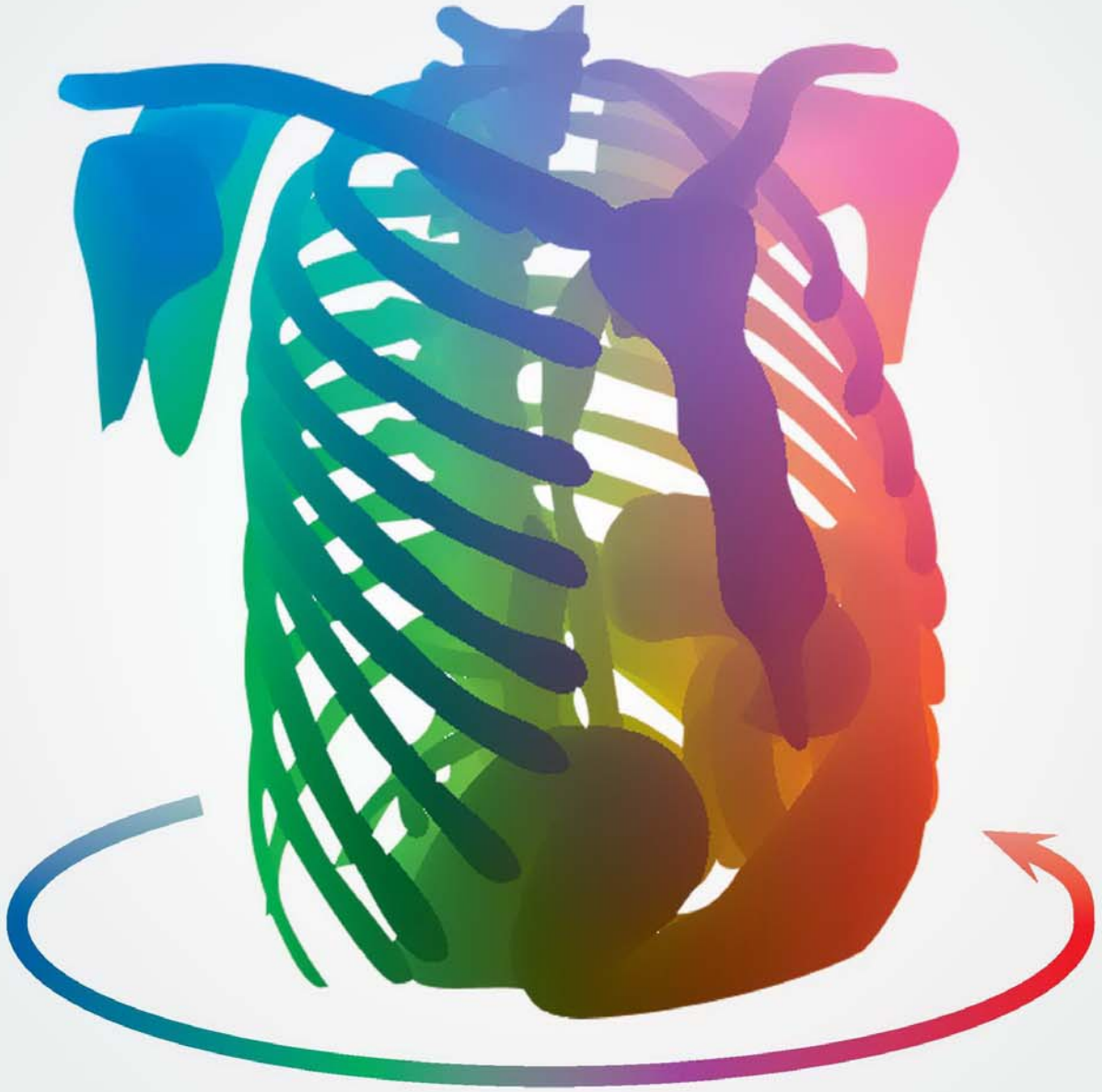
JIG
JOURNAL OF IMAGE AND GRAPHICS
主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2015
02
VOL.20
ISSN1006-8961
CN11-3758/TB



等值面光线跟踪 *P193*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第20卷第2期（总第226期）
2015年2月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 李小文
编辑出版 《中国图象图形学报》编辑出版委员会
邮政信箱 北京9718信箱
邮　　编 100101
电子信箱 jig@irsa.ac.cn
电　　话 010-64807995
网　　址 www.cjig.cn

广告经营许可证 京朝工商广字第0361号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics
Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief LI Xiaowen
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
P.O.Box 9718, Beijing, P.R.China
Zip code 100101
E-mail jig@radi.ac.cn
Telephone 010-64807995
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ
国外发行代号 M1406
国内邮发代号 82-831
国内定价 50.00元

2015年2月 第20卷 第2期（总第226期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

多尺度活动网格在云场景仿真中的应用(第0202页)

自然环境视频中基于显著鲁棒轨迹的行为识别(第0245页)

虚拟手抓持力觉生成算法真实性的评价(第0280页)

Volume 20, Number 2
Published February 16, 2015

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Multi-scale moving grids and its application in clouds scene simulation(P0202)

Action recognition from unconstrained videos via salient and robust trajectory (P0245)

Validation the Authenticity of Virtual fingers Grasping Haptic Algorithm(P0280)

中图法分类号：TP301.6　　**文献标识码**：A　　**文章编号**：1006-8961(2015)02-0211-08

论文引用格式：Yang P, Xie L, Zhang Y N. Survey on unsupervised spoken term detection for low-resource languages[J]. Journal of Image and Graphics, 2015, 20(2): 0211-0218. [杨鹏, 谢磊, 张艳宁. 低资源语言的无监督语音关键词检测技术综述[J]. 中国图象图形学报, 2015, 20(2): 0211-0218.] [DOI: 10.11834/jig.20150207]

低资源语言的无监督语音关键词检测技术综述

杨鹏，谢磊，张艳宁

西北工业大学计算机学院陕西省语音与图像信息处理重点实验室，西安　710129

摘　要：目的　低资源(low-resource)语言的无监督的关键词检测技术近年来引起了广泛的研究兴趣。低资源语言由于缺乏足够的标注数据及相关的专家知识，使得传统的基于大词汇量语音识别系统的关键词检测技术无法使用。近年来，研究者试图寻找一种无监督的技术来完成针对低资源语言的语音关键词检测。**方法**　首先阐述了该技术目前面临的问题与挑战，然后介绍了该技术使用的主流的基于动态时间规整的算法框架，并从特征表示、模板匹配方法、效率提升等几个重要方面介绍了近几年来主要的研究成果，最后介绍了该任务常用的系统评价标准及目前所能达到的水平，讨论了未来可能的研究方向。**结果**　该任务的研究目前取得了很多成果，但仍处于实验室阶段，多系统融合策略导致系统庞大，而且目前还没有好的进行索引的方法，导致检测时间过长，对于低资源语音的关键词检测技术，还有很多研究工作要做。**结论**　期望通过对目前低资源语言的无监督的关键词检测技术做出一个全面的综述，从而给研究者的工作带来便利。

关键词：语音关键词检测；低资源；动态时间规整

Survey on unsupervised spoken term detection for low-resource languages

Yang Peng, Xie Lei, Zhang Yanning

Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xi'an　710129, *China*

Abstract: Objective Query-by-example spoken term detection for low-resource languages has recently drawn considerable research interest. For low-resource languages that lack sufficient annotated data and related expert knowledge, spoken term detection techniques based on traditional large vocabulary speech recognition cannot be directly used. Researchers have recently attempted to determine an unsupervised technique to perform this task for low-resource languages. **Method** In this study, we first present the challenges confronting this task. We then introduce the algorithm framework based on dynamic time warping (DTW) commonly used in this task. We finally present the recent research devoted to feature representation, template matching, speed-up, and other related topics. **Result** Although the research of this technique on low-resource language has got much progress, there are not real-life applications. Some unified feature representation and indexing method must be proposed to attain both good effectiveness and efficiency. **Conclusion** We present the commonly used performance evaluation standards. The conclusion of our investigation is presented, and possible future research directions are discussed.

Key words: spoken term detection; low-resource; dynamic time warping

收稿日期：2014-07-25；**修回日期**：2014-11-06

基金项目：国家自然科学基金项目(61175018)；霍英东青年教师基础研究基金项目(131059)

第一作者简介：杨鹏(1989—)，男，现为西北工业大学计算机学院计算机应用技术专业硕士研究生，主要研究领域为语音模式发现、语音关键词检测。E-mail: pengyang@nwpu-aslp.org

Supported by: National Natural Science Foundation of China (61175018); The Fok Ying-Tong Education Foundation, China (131059)

0 引 言

语音关键词检测(STD)任务是指从给定的语料库中检测出指定语音关键词的所有出现[1]。目前多数的语音关键词检测技术是基于大词汇量连续语音识别系统(LVCSR)的方法。该方法以语音关键词文本作为输入,通过对语料库进行语音识别,将其转化为识别词网格表示,然后再在该词网格上进行关键词匹配[1-2]。此类方法虽然能够达到较好的检测效果,但是训练一个大词汇量语音识别系统需要大量的标注数据资源。特别是针对一种新的语言,在没有任何语言学专家知识和标注数据的情况下,用该类方法是无法实现关键词检测的。此外,目前大多数的语言都是低资源(low-resource)的,即缺乏语言学专家知识和足够的具有抄本标注的语音数据,并不具备训练一个鲁棒的语音识别系统的条件[3]。另一方面,大词汇量语音识别系统普遍面临着词典未登录词汇(OOV)问题的困扰。在语音关键词检测的任务中,用户输入的语音关键词往往是一些人名、地名、事件名等,这些专有名词很可能就是OOV词汇,从而使此类关键词检测效果大打折扣。

近几年来,无监督语音关键词检测(QbE-STD)技术开始引起广泛研究。该类技术以一个音频样本作为输入,通过与语料库中的语音音频文件直接进行声学相似度的计算,从给定语料库中返回含有该语音关键词的所有文档。该框架无需训练一个语音识别系统,从而避免了OOV问题的困扰,同时也规避了低资源语音不具备训练语音识别系统条件的问题。由欧美多所大学和研究机构组织的MediaEval多媒体国际评测从2011年起加入了语音关键词检测任务——Spoken Web Search (SWS),该任务要求研究者使用关键词的语音样本作为输入,在语音语料中检测这些关键词的出现。该评测任务提供了时长为20 h左右的语料库和个数分别为500左右的发展集语音关键词(development set:用以调参)和评测集语音关键词(evaluation set:用以系统最终评测)。该任务由多种低资源语言的数据组成(例如某种非洲语言),没有任何标注抄本和语言本身信息。此外,语音数据完全在日常条件下进行录制,部分语音带有明显的口音和噪声,增加了挑战性。该任务引起众多研究机构的兴趣,例如约翰霍普金斯大学、卡耐基梅隆大学,每年都会有一批高质量的论文发表在语音领域国际顶级会议ICASSP、Interspeech上[4-9]。

无监督的语音关键词检测技术尚属起步阶段,虽然在检测效果和效率上都无法和传统基于LVCSR的方法抗衡,然而其"低资源"、"轻量级"的优点,吸引了众多研究者开始投入研究工作。目前主流的算法框架是基于动态时间规整(DTW)的模板匹配算法的各种变体。当前的研究主要集中在对说话人和环境鲁棒的特征表示、模板匹配方法和效率提升上。本文首先对目前主流的无监督语音关键词检测算法框架进行介绍,然后从语音特征表示、匹配方法、效率提升3个方面对近年来的研究成果加以综述,最后将介绍目前该任务使用的评测标准和SWS国际评测中的无监督关键词检测任务中最近一次评测结果中性能最好的系统。本文期望通过对目前低资源语言的无监督的关键词检测技术做一个全面的综述,从而给研究者的工作带来便利。

1 算法框架简介

无监督的语音关键词检测现在主流的算法框架是基于动态时间规整(DTW)的模板匹配算法[10]。该框架首先使用合适的特征表示语音关键词和语料库中的所有语料,然后通过动态时间规整算法的某种变体找出给定语音关键词在每个语音文件中的出现位置,并给每个语音文件打分,定义其含有指定语音关键词的可能性,然后再依此分数按一定顺序返回结果。

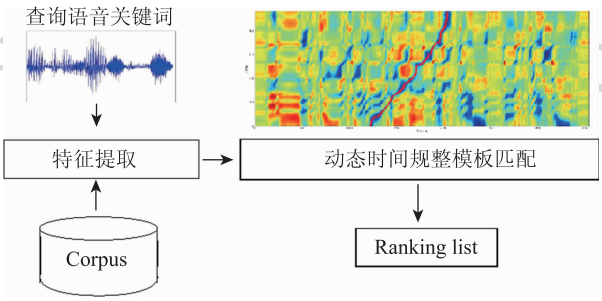


图1 基于动态时间规整的语音关键词检测的算法框架

Fig. 1 The DTW-based spoken term detection framework

如图1所示,该算法框架主要有两个模块组成,一个是特征提取模块,另一个是模板匹配模块。特

征提取模块需要将原始的波形文件表示的语音数据以一种更好的反映其发音信息的特征表示出来。而模板匹配模块的职责是准确并且快速地找到与语音关键词声学上“最像”的区域，并以打分的形式加以量化。该模块首先建立语音关键词的特征向量序列与语料库中音频文件的特征向量序列之间的距离矩阵，然后通过动态时间规整算法找出关键词最有可能出现的位置。图 1 中距离矩阵的 dot-plot 图中的斜线所代表的起始点即算法找到的语音关键词出现的位置。近些年来基于此框架的研究成果众多，下面将从特征表示、模板匹配方法、效率提升 3 个方面加以介绍。

2　特征表示

众所周知，短时谱特征，例如美尔域倒谱系数（MFCC），感知线性预测编码（PLP），是目前语音识别领域广泛使用的声学特征。短时谱特征较好地描述了人在发音时的声道状态信息或人耳对语音的感知信息，但同时也包含了说话人信息、环境噪音信息和传输信道信息，这些“个性”信息降低了语音识别器的性能。因此，在无监督的关键词检测任务上，原始的短时谱特征同样效果不佳。

特征表示的研究目标就是希望从原始谱特征中保留发音时音素类别判别信息，而去除说话人、环境噪声以及传输信道等信息。

2.1　后验特征

目前主流的特征提取方法是使用后验特征（Posteriorgram）。后验特征，即给定一个语音特征向量 $\boldsymbol{o}$，该特征向量在预先定义好的 k 个类 $\{C_1, C_2, \cdots, C_k\}$ 上的后验概率分布为

$$\boldsymbol{PG}_o = (p(C_1 \mid \boldsymbol{o}), \cdots, p(C_k \mid \boldsymbol{o})) \qquad (1)$$

式中，$p(C_i \mid \boldsymbol{o})$ 是特征向量 $\boldsymbol{o}$ 第 i 个类上的后验概率。

文献[11]中使用了音素后验特征，具体做法是使用训练好的音素识别器对语音数据进行识别得到识别网格，然后再使用识别网格得到每一帧语音数据的音素级的后验概率。目前音素后验特征是 QbE-STD 框架下效果最好的特征表示方法。但这种方法实际上是一种有监督地学习方法，需要准备具有音素标注抄本的训练数据以及进行音素模型训练。针对低资源语音数据，根本不具备训练相应音素识别器的条件，音素后验特征通常使用一种资源丰富的语言的语料训练一个音素识别器，然后的去解码出低资源语音的后验特征，然后再进行后续的算法步骤。虽然不同语种之间的语音发音有较高的相关性，但是这种“不匹配”的方法，关键词检测效果会大打折扣。实际系统中，经常是训练多种“高”资源语言语音的音素识别器（如英语、法语），然后将多个系统进行融合，具体策略有文献[7]中后端的分数融合策略和文献[8]中将多个距离矩阵进行融合的策略。

图 2 是语音片段“美国总统克林顿”的音素后验特征向量序列随时间变化的图[18]。图 2 采用 25 ms 帧长，10 ms 帧移的设置。纵轴上的英文字符串，分别对应 67 个音素，即式(1)中的 k 等于 67，所以图 2 中的音素后验特征向量是一个 67 维的向量。特定语音帧属于某个音素的后验概率越大则越黑，越小则越白。从图 2 可以看出，文献[11]通过使用音素标注数据有监督地训练一个音素识别器，从原始谱特征中将音素“类别”信息提取了出来，所以音素后验特征在一定程度上抹掉了原始谱特征含有的关于说话人，以及传输信道的信息，从而使得模板匹配的效果得到提升。

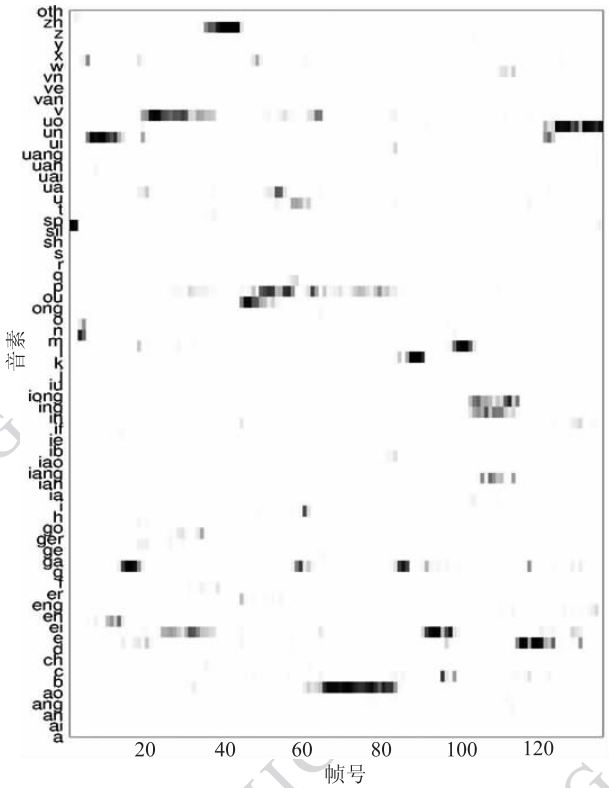


图 2　某语音片段的后验特征向量序列随时间变化图[18]

Fig. 2　The posteriorgram dot-plot of one speech segment

为了规避音素后验特征需要标注数据进行训练的缺点,文献[12]提出了高斯后验特征,具体做法是:使用语料库中部分或者所有数据无监督地训练一个高斯混合模型(GMM),然后计算每一帧语音数据在每个高斯混合分量上的后验概率,从而得到帧级的后验特征。高斯后验特征在实践中的效果,与上述提到的"不匹配"地使用的音素后验特征效果相当[9]。文献[13]使用了深度玻尔兹曼机(DBM)来获得后验特征,精调(fine tuning)阶段使用了GMM模型[12]给每帧语音打标签。这种无监督地获取后验特征的方法取得了较高斯后验特征更好的效果,文献[13]将此归功于深度学习带来的效果。

另外一种获得后验特征的方法是:通过无监督的方式自动发现语音中的子词模式,比如类似音素单元[11],然后对这些子词模式进行声学建模,从而无监督地得到了一个子词识别器,然后使用与文献[11]中相同的方法获得后验特征。文献[14]将此无监督地建立子词识别器的过程分为3步:1)分割,将连续语音信号分割成离散的、非重叠的片段,使得每个语音片段对应于某个音素的一次发音实现。目前最成熟稳定方法是文献[15]中提出的基于信息论中率失真(rate distortion)概念的分割方法。2)聚类,对于上一步得到的语音片段集合,采用某种聚类方法,比如,k均值聚类,将声学相似的语音片段聚为一类,代表某个子词模式的多个实现。3)建模,对每一类子词模式进行声学建模,一般使用隐马尔可夫模型(HMM)。

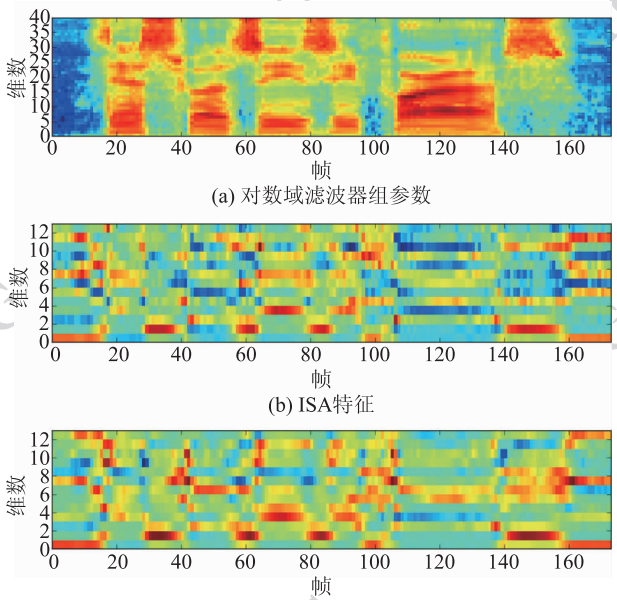
文献[9]使用上述策略,同时将声道归一化(VTLN)和CMLLR(constrained maximum likelihood linear regression)[16]两种说话人归一化技术引入在了声学建模的过程中,最后得到的后验特征在关键词检测效果上比高斯后验特征[12]有明显提高。文献[17]则使用非参贝叶斯方法——狄利克雷过程混合模型(DPM),将语音片段的分割边界设为了模型参数,使用3个状态HMM模型对每个语音片段进行建模,使用Gibbs采样器在最大似然的意义下来估计所有的模型参数。该方法的特点是:将分割、聚类、建模3个问题看成一个整体来进行建模,而不是采用文献[14]中提出的"三步走"策略。在"三步走"策略中,每一步的解决方案是割裂的,不考虑其对后续步骤的影响。文献[17]的方法在语音关键词检测任务中的效果要优于高斯后验特征[12]和深度玻尔兹曼机的后验特征[13]。

2.2 声学特征

除MFCC、PLP等谱特征被广泛采用之外,研究者也在不断尝试新的更具有发音类别判别能力的新特征。研究发现,由于人类发声器官运动的限制,语音被限制在一个高维空间的低维流形上。基于这一发现,文献[19]提出了一种流型学习算法ISA(intrinsic spectral analysis)。该算法通过无监督学习,得到一组非线性的内在基函数,将高维的原始语音谱特征变换到低维的流型上。实验结果证明,该特征与使用传统的离散余弦变换(DCT)得到的MFCC特征相比,能更好地表征语音的声学区分性,在孤立词匹配任务上ISA特征取得了良好的效果。

近来,受到上述ISA特征的启发,同时考虑到语音中普遍存在的协同发音现象,即任何语音帧都会受到上下文语音帧的影响,把每个语音帧用前后多帧拼接起来形成超向量表示,获得了更好地反映原始语音数据流型结构的内在基函数。将此"瞬时上下文"ISA特征应用到了语音关键词检测的任务上[20],在TIMIT语料库上的实验表明,此种改进方法得到的低维特征取得了相比于原始ISA特征[19]更好的结果。

图3为TIMIT库中截取单词"organization"的3种特征表示的dot-plot图。从图3中可以看出,ISA[18]



(a) 对数域滤波器组参数

(b) ISA特征

(c) 基于上下文信息的ISA特征[20]

图3 英文单词"organization"的3种特征表示的dot-plot图

Fig. 3 The three types of feature representation's dot-plot of English word "organization"

通过使用一组内在基函数将 40 维的原始谱特征(log-scale filter-bank parameters)变换到 13 维。这 13 维特征随时间的变化与单词“organization”发音变化有很强的相关性。文献[20]通过在流型学习的过程中引入上下文信息,获得了更好的内在基函数的估计,从而在图 3(c)中,13 维特征随时间的变化显得更加“平滑”和“连续”,不似图 3(b)那样存在明显的“噪点”。文献[20]中的关键词检测实验也证明了这种改进的有效性。

3　模板匹配方法

无监督的关键词检测任务可以看成是一个信息检索任务,因此该系统需要定义语料库中每一个语音文档含有指定查询语音关键词(spoken query)的可能性,或者使用信息检索领域的术语来讲,即相关性(Relevance)。DTW 算法的各种变体被广泛地用来定义查询语音关键词(spoken query)和特定语音文档的相关性。从最初的 segmental-DTW[8-9,12] 到 Subsequence DTW[10],再到最近提出的 segmental local-normalized DTW[6,20,22],算法从计算效率和匹配效果上都有明显提升。

Segmental-DTW[12] 匹配路径是在固定大小的窗口进行,通过以一定步长移动窗口,将匹配路径打分最高的窗口对应的区域作为检测结果。这种滑动窗口的策略一方面导致了多次在窗口里进行 DTW 计算,效率低;另一方面待检测的区域只能和查询语音关键词在时间上等长,这显然与多次发音在时间上存在长短变化的事实不符。

Subsequence DTW[21] 算法通过对 DTW 算法初始条件的改进,即允许待查询语音片段的任何一帧作为 DTW 的起点,通过一次递归填写矩阵的方法,找到最佳匹配区域,提升了效率。另外 Subsequence DTW 检测到的发音相似区域的长度,是由数据本身决定的,这个与语音数据时域变化性符合。

Segmental local-normalized DTW[6,20,22] 相比于 Subsequence DTW,其最大的特点在于加入了 local-normalized 的概念,使得 DTW 算法的距离积累矩阵里的值是原本的积累距离值除以其当前匹配路径的长度。这种在动态规划的过程中就开始进行长度归一化的方法,增强了算法在模式匹配上的鲁棒性。

目前类 Segmental local-normalized DTW 算法已经成为了基于动态时间规整算法框架下模板匹配模块最常用的算法。下面将此算法做一较为详细的介绍。语音关键词 $\boldsymbol{q}=\{\boldsymbol{q}_1,\cdots,\boldsymbol{q}_n\}$,其中 n 为语音关键词的帧数。待检测的语音文档为 $\boldsymbol{t}=\{\boldsymbol{t}_1,\cdots,\boldsymbol{t}_m\}$,其中 m 为待检测语音文档的帧数。定义 $\boldsymbol{d}(i,j)$ 表示 $\boldsymbol{q}_i$ 和 $\boldsymbol{t}_j$ 之间的距离。然后计算一个 $n\times m$ 的距离矩阵 $\boldsymbol{cost}$,其中 $cost(i,j)=a(i,j)/l(i,j)$,$a(i,j)$ 代表了从某个起点到达(i,j)所经历的路径积累距离,而 $l(i,j)$表示从某个起点开始到达(i,j)所经过的路径长度。所以 $cost(i,j)$代表了从某个起点到达点(i,j)的平均积累距离。该问题可以使用动态规划算法解决。$a(i,j)$和 $l(i,j)$初始化为

$$a(1,j)=d(1,j),l(1,j)=1;$$

$$a(i,1)=\sum_{k=1}^{i}d(k,1)/l(i,1),l(i,1)=i;$$

对于其他(i,j),首先从$(i,j-1)$,$(i-1,j)$和$(i-1,j-1)$3 个前继点中找一个点$(\hat{r},\hat{c})$,使得表达式

$$\frac{a(\hat{r},\hat{c})+d(i,j)}{l(\hat{r},\hat{c})+1}$$

最小。语音文档 $\boldsymbol{t}$ 的分数被定义为 min $cost(n,j)$。最后所有语音文档以此分数从小到大排序,即为返回结果。

4　效率提升

目前主流的无监督语音关键词检测技术都是基于动态时间规整算法的,而动态时间规整算法的时间复杂度为 $O(mn)$,对于大数据量的关键词检测任务来说,时间代价依然过大。现在主要的加速方法分为两种:一种是通过启发式算法,较快定位语料库中与查询关键词发音相似的候选区域,然后再使用 DTW 算法计算该区域的相似度打分。另一种方法是对语料库离线建立索引,依此来获得在线检索时的速度提升。

4.1　在线方法

麻省理工学院的 Zhang 等人提出了基于 DTW 下界(lower-bound)的加速方法[23-24] 以及将下界算法采用 GPU 进行再次加速[25]。最近文献[26]提出了一个“更紧”的下界,采用该下界可以得到更快的加速效果,而且可以以牺牲空间为代价换取更大的加速效果。文献[23-26]的这种基于 DTW 下界算法的加速策略都是针对 segmental DTW[13] 算法提出

的。这类算法首先给语料库中所有与查询关键词等长的语音片段计算一个 DTW 打分的下界估计,然后通过该下界估计对所有语音片段进行排序,形成一个优先队列,越“像”的语音片段越早地出队列,使用传统的 DTW 算法进行相似度打分,当只需要返回前 k 个最好的结果时,可以利用当前优先队列首部语音片段的下界估计和当前返回结果中“第 k 好”的语音片段的 DTW 打分进行比较,如果下界估计已然较大,则当前返回队列中的 k 个语音片段即是语料库中最相似的前 k 个“语音片段”,从而停止搜索。这种方法利用一种可以快速计算的 DTW 打分的下界估计,来避免 segmental DTW 算法需要与语料库中所有语音片段进行 DTW 算法的特点,减少了 DTW 算法执行的次数,从而提高了效率。而最近,文献[27]指出,segmental local-normalized DTW 在算法复杂度上和上面提到的基于下界的加速算法是相当的,该算法并不需要计算下界[22],也不需要牺牲空间[26],通过在算法中放松传统 DTW 算法中的对于起点的限制,允许匹配路径起点开始于待查询语音文件的任何一帧,避免 segmental DTW[12]的滑动窗口策略导致的多次 DTW 计算,提高了效率,同时提升匹配效果。

4.2 离线方法

离线方法是指,通过对语料库离线建立索引,依此来获得在线检索时的速度提升。文献[28]中使用了一种“粗搜”加“细搜”的两步搜索方法。首先,使用层次聚类方法,对语音文件进行分割,对每个分割片段使用其谱特征向量的均值表示,此时就形成了一个“降采样”版的文档。在这个“降采样”的文档上进行粗搜,称为 segment-based DTW,找出大致位置。而后,使用传统的 DTW 算法在该位置上进行细搜与确认。这种方法可以看成是使用离线(off-line)的预处理对语料库建立索引,然后以离线的时间消耗来换取在线(on-line)检测时的速度提升。文献[29]使用层次 k-means 算法将语料库中的语音数据帧进行聚类,形成一个树状结构,语音关键词中的特定帧可以通过此树结构迅速找到与其相似的语音帧,从而加速了文献[29]中使用的检索算法。文献[30]对语料库中每个语音文档依某个特定的时长,分成很多片段,然后以这些片段为结点建立一个图,使用贪心搜索(greedy search)和广度优先搜索(breadth-first search),对于特定语音关键词在这个图上进行检索。因此,该图就是一种索引。实际应用中,由于语音关键词的长度变化较大,所以在将语料库中的语音文档分成很多片段的时候,会使用多个时长,所以就建立多个图索引,当特定的语音关键词被送到检索系统时,选择与该语音关键词的时长最为接近的图索引进行检索。

5 系统评测标准与语料库

对于检测效率的评价,文献[28]提议使用平均查询时间和语料库总时长的比率来衡量。但是,目前大多数论文只是采用了经验性的绝对时间来对比各种算法效率的优劣。

对于效果的评价,目前主要使用 MAP(mean average precision)和 ATWV(actual term weighted value)来衡量。对于某个语音文档,若该文档含有对应的语音关键词,则称为命中,否则称为漏检。MAP 是信息检索任务中常用的一种评价标准,是对精准度的一个平均的测度。MAP 越高,表示靠前的返回结果中命中得多,相反则漏检得越多。MAP 的计算可以使用 trec_eval 工具(http://trec.nist.gov/trec_eval/)来完成。ATWV 是关键词检测任务中常用的评测标准。NIST 组织的传统的基于识别词网格的关键词检测任务,一直使用该标准来衡量系统的性能。ATWV 是对精准度(precision)和召回度(recall)的一个折中,可以采用 Mediaeval 2013 workshop 发布的工具(http://www.multimediaeval.org/mediaeval2013/)来计算。

在评测数据库方面,经典的 TIMIT 语料库[31]被经常用来验证各种算法的效果。除此之外,目前 Mediaeval 2013 的无监督关键词检测任务 SWS,其数据包已经公布[32]。该数据包中包含有 20 h 的语料、500 个发展集查询语音关键词、500 个评测集查询语音关键词、评测脚本及相关的参考资料。该数据包的公布给无监督关键词检测任务的研究者带来了便利,研究者可以方便地使用这些数据评测自己的系统,并与其他系统做比较。

Mediaeval SWS 2013 评测中,文献[6]中描述的系统取得了最好的结果,其 ATWV 达到了最高,0.444。该系统主要思路在于:利用多个“高”资源语言的音素识别器得到评测数据的多种后验特征表示,然后利用不同后验特征之间的互补信息,将不同

特征表示下的系统输出结果进行融合，从而取得较只使用单一特征更好的结果。值得注意的是，简单的特征拼接带来的效果提升，并没有后端的系统分数融合[7]的效果好。

6 结　语

无监督的语音关键词检测技术，旨在少量借助，甚至不借助任何专家标注信息，来完成低资源语音数据的关键词检测。这种技术框架避免了传统的基于大词汇量语音识别系统的 OOV 问题，同时由于该技术的语言独立性，使得它可以应用于任何低资源的语言，不受标注数据昂贵稀缺的限制。相对于传统的基于识别词网格的方法，这种无监督技术具有“轻量级”的优点，但由于该技术的研究刚刚起步，检测效果和检测效率都还有很大的提升空间。

未来潜在的研究方向有：1）利用深度学习来提取更好地表征语音声学变化的特征。2）尝试 GPU 及分布式计算等最新技术，来提高无监督关键词检测系统的运行效率，使其适用于真实的大数据的任务。3）尝试使用无监督的声学建模技术，来建立语音识别器，从而利用成熟的基于识别词网格的技术来完成关键词检测。

参考文献（References）

[1] Chelba C, Hazen T J, Sarac, lar M. Retrieval and browsing of spoken content[J]. IEEE Signal Process. Mag., 2008, 25(3): 39-49.

[2] Saraclar M, Sproat R. Lattice-based search for spoken utterance retrieval[C]//Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics. North American: Association for Computational Linguistics, 2004: 129-136.

[3] Glass J. Towards unsupervised speech processing[C]// Proceedings of the 11th International Conference on Information Sciences, Signal Processing and their Applications. Montreal, USA: IEEE, 2012: 1-4.

[4] Metze F, Rajput N, Anguera X, et al. The spoken WEB search task at mediaeval 2011[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012: 5165-5168.

[5] Metze F, Anguera X, Barnard E, et al. The spoken web search task at mediaEval 2012[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE, 2013: 8121-8125.

[6] Rodriguez-Fuentes L J, Varona A, Penagarikano M, et al. High-performance query-by-example spoken term detection on the SWS 2013 evaluation[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2014: 7869-7873.

[7] Abad A, Rodriguez-Fuentes L J, Pena-garikano M, et al. On the calibration and fusion of heterogeneous spoken term detection systems[C]//Proceedings of the 17th Annual Conference of the International Speech Communication Association. Lyon, France: ISCA, 2013.

[8] Wang H, Lee T, Leung C C, et al. Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE, 2013: 8545-8549.

[9] Wang H, Leung C C, Lee T, et al. An acoustic segment modeling approach to query-by-example spoken term detection[C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012: 5157-5160.

[10] Müller M. Dynamic time warping[J]. Information Retrieval for Music and Motion, 2007: 69-84.

[11] Hazen T J, Shen W, White C. Query-by-example spoken term detection using phonetic posteriorgram templates[C]// Proceedings of the IEEE Automatic Speech Recognition & Understanding. Merona, Italy: IEEE, 2009: 421-426.

[12] Zhang Y, Glass J R. Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams[C]//Proceedings of the IEEE International Workshop on Automatic Speech Recognition & Understanding. Merano, Italy: IEEE, 2009: 398-403.

[13] Zhang Y, Salakhutdinov R, Chang H A, et al. Resource configurable spoken query detection using deep Boltzmann machines[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012: 5161-5164.

[14] Wilpon J, Juang B, Rabiner L. An investigation on the use of acoustic sub-word units for automatic speech recognition[C]// Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Texas, USA: IEEE, 1987: 821-824.

[15] Qiao Y, Shimomura N, Minematsu N. Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, USA: IEEE, 2008: 3989-3992.

[16] Gales M J F. Maximum likelihood linear transformations for HMM-based speech recognition[J]. Computer Speech & Lan-

guage, 1998, 12(2): 75-98.

[17] Lee C, Glass J. A nonparametric bayesian approach to acoustic model discovery[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju island, Korea:IEEE, 2012: 40-49.

[18] Yang P, Xie L. Mandarin speech pattern discovery using segmental dynamic time warping and posteriorgram features [J]. Journal of Tsinghua University (Sci. & Tech.), 2013, 53(6): 903-907. [杨鹏,谢磊. 基于动态时间规整和后验特征的中文语音模式发现[J]. 清华大学学报: 自然科学版, 2013, 53(6):903-907].

[19] Jansen A, Niyogi P. Intrinsic spectral analysis [J]. IEEE Transaction on Signal Processing, 2013, 61(7):1698-1710.

[20] Yang P, Xie L, Leung C C, et al. Intrinsic spectral analysis based on temporal context features for query by example spoken term detection[C]// Proceedings of the 18th Annual Conference of the International Speech Communication Association. Singapore:ISCA, 2014, 1722-1726.

[21] Anguera X, Ferrarons M. Memory efficient subsequence DTW for query-by-example spoken term detection[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. California, USA:IEEE, 2013: 1-6.

[22] Muscariello A, Gravier G, Bimbot F. Audio keyword extraction by unsupervised word discovery [C]//Proceedings of the 10th Annual Conference of the International Speech Communication Association. Brighton, UK:IEEE, 2009: 2843-2846.

[23] Zhang Y, Glass J R. An inner-product lower-bound estimate for dynamic time warping [C]//Proceedings of the IEEE International Conference on Acoustics, Seech and Signal Processing. Prague, Czech:IEEE, 2011: 5660-5663.

[24] Zhang Y, Glass J R. A piecewise aggregate approximation lower-bound estimate for posteriorgram-based dynamic time warping [C]//Proceedings of the 15th Annual Conference of the International Speech Communication Association. Singapore: IEEE, 2011: 1909-1912.

[25] Zhang Y, Adl K, Glass J. Fast spoken query detection using lower-bound dynamic time warping on graphical processing units [C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan:IEEE, 2012: 5173-5176.

[26] Yang P, Xie L, Luan Q, et al. A tighter lower bound estimate for dynamic time warping[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada:IEEE, 2013: 8525-8529.

[27] Mantena G, Achanta S, Prahallad K. Query-by-Example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping [J]. IEEE Transaction on Audio, Speech, and Language Processing, 2013, 22(5): 946-955.

[28] Chan C, Lee L. Integrating frame-based and segment-based dynamic time warping for unsupervised spoken term detection with spoken queries[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Prague, Czech:IEEE, 2011: 5652-5655.

[29] Mantena G, Anguera X. Speed improvements to information retrieval-based dynamic time warping using hierarchical k-means clustering[C]//Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing. Vancouver, Canada:IEEE, 2013: 8515-8519.

[30] Aoyama K, Ogawa A, Hattori T, et al. Graph index based query-by-example search on a large speech data set [C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada:IEEE, 2013: 8520-8524.

[31] Garofolo J S, Lamel L F, Fisher W M, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium [EB/OL] (1993) [2014-6-21]. https://catalog.ldc.upenn.edu/ LDC93S1.

[32] SWS 2013 Multilingual Database for Query-by-Example Keyword Spotting [EB/OL]. (2013-10-12) [2014-6-21]. http://speech.fit.vutbr.cz/software/sws-2013-multilingual-database-query-by-example-keyword-spotting, 2014,6,21.