基于特征空间轨迹信息的语音关键词检测方法

田颖慧,贺前华,郑若伟,危 卓,李艳雄
(华南理工大学,广东广州 510641)

摘 要: 当前语音关键词检测的主流技术为深度学习,需要大规模标注样本进行训练,难以应用于更普遍的低资源场景. 本文提出一种基于音频特征空间轨迹信息的低资源语音关键词检测方法,该方法基于"词是由更小语言单元(音节、音素)的结构化组成,以及语言单元声学特征具有稳定性(统计意义)"的事实,结合物理几何空间定位的原理,构建语音关键词的特征空间表达、时序信息表达和局部区分信息知识. 语音关键词检测时,依据语音段的特征空间轨迹信息分层次进行判决,实现了模式信息与统计信息的综合应用. 其中语音特征空间是利用丰富的无标注语音样本构建音频特征空间的标识子表达,而语音关键词的特征空间轨迹信息利用少量关键词语音样本构建. 多个实验结果表明,本文算法在低资源时(100个样本以下),相比HMM和CRNN有显著优势,10个训练样本时,相比HMM,FRR绝对下降了20.5%,FAR绝对下降了8.7 FP/h;而在训练样本量较充分(300个样本及以上)时,与CRNN有大致相当的性能.

关键词: 语音关键词检测;音频特征空间;特征空间轨迹信息;低资源
基金项目: 广东省自然科学基金(No.2022A1515011687);国家自然科学基金(No.61571192)
中图分类号: TP391.4;TP391.9 **文献标识码:** A **文章编号:** 0372-2112(2023)10-2915-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220289

Spoken Term Detection Based on Feature Space Trajectory Information

TIAN Yin-hui, HE Qian-hua, ZHENG Ruo-wei, WEI Zhuo, LI Yan-xiong
(*South China University of Technology*, *Guangzhou*, *Guangdong* 510641, *China*)

Abstract: The current technique of spoken term detection is dominated by deep learning, which requires large annotated data for training, and is difficult to be applied in limited-data scenarios. In this paper, a feature trajectory based method of spoken term detection is proposed for limited-data scenarios. The method originated from the fact that a word is a structured organization of small units such as syllable or phoneme and any language unit has steady statistical audio feature, based on the principle of physical location, feature distribution, temporal information of keywords, and local distinguishing information are constructed with speech examples. Spoken keywords are searched with the feature trajectory information of the detected speech segment in hierarchical decision strategy. The method works on a audio feature space defined by a identifier set trained with a large unlabeled speech dataset. Several experimental results show that the proposed method is evidently superior to HMM and CRNN when the training samples is less than 100. For example, when 10 samples are used for training, FRR and FAR of the propose method are absolutely decreased by 20.5% and 8.7 FP/hour respectively compared with HMM-based system. On the other hand, the proposed method achieved the comparable performance v.s. CRNN-based system when the training samples is more than 300.

Key words: spoken term detection; audio feature space; feature space trajectory information; limited-data source
Foundation Item(s): Guangdong Natural Science Foundation (No. 2022A1515011687); National Nature Science Foundation of China (No.61571192)

1 引言

语音关键词检测(Spoken Term Detection,STD)是指从连续语音流中检测预先定义的关键词,已被广泛应用于信息检索[1,2]、音频监控[3]及语音助手[4,5]等多个领域. 其技术路线根据数据资源可以分为两大类型. 一种是应用于富资源数据的关键词检测方法,如基于大词汇量连续语音识别(Large Vocabulary Continuous Speech Recognition,LVCSR)、声学模型(Acoustic Model,AM)的方法;另一种是应用于低资源甚至零资源数据的关键词检测方法,如按样例查询(Query by Ex-

收稿日期:2022-03-22;修回日期:2022-08-26;责任编辑:朱梅玉

ample,QbyE).

基于LVCSR[6,7]的方法通过语音识别系统生成词格,对词格进行搜索来预测关键词是否存在.虽然该方法可以灵活地改变关键词,但是关键词检测的性能依赖语音识别系统的性能.语音识别系统需要大量的标注数据[8],对于无标注数据的小语种或方言来说难以适用.

基于AM[9]的方法,例如传统的高斯混合模型-隐马尔可夫模型(Gaussian Mixture Model-Hidden Markov Model,GMM-HMM)分别为关键词和非关键词建模,由于少量非关键词模型很难充分表达所有的非关键词,因此通常具有较高的虚警率,同时模型训练依赖数千小时的训练数据[10,11].随着深度学习的发展,基于深度神经网络-隐马尔可夫模型(Deep Neural Network-Hidden Markov Model,DNN-HMM)的方法,如文献[12],HMM解码器结合DNN的预测计算得分,但DNN的训练忽略HMM参数,存在任务目标和训练目标失配的问题.文献[13]虽然设计端到端的损失函数缓解了DNN单独训练带来的负面影响,但模型训练需要50万条包含关键词的语音样本.基于卷积神经网络(Convolutional Neural Network,CNN)[4]和卷积循环网络(Convolutional Recurrent Neural Network,CRNN)[14]的方法相比于HMM,关键词检测的性能得到显著提升,但是模型训练依赖帧级别的标签,需要对语音和文本做预对齐处理,产生大量的标注成本.端到端的方法由于不需要对语音和文本做预对齐处理而受到广泛关注,如文献[15,16],省去对数据做预对齐的成本.但是模型训练依然需要数百或数千小时的具有转录文本的数据,在低数据资源下并不适用.

基于模板匹配的方法,如QbyE[17],直接利用样例基于动态时间规整(Dynamic Time Warping,DTW)对音频样本进行匹配搜索,不需要标注数据训练模型,可用于低数据资源的场景.但是QbyE只根据样例与测试样本的距离进行判决,缺乏对不同类别信息的利用.DTW利用动态规划(Dynamic Programming,DP)思想搜索累积距离最小的路径,着重考虑时序序列基于数值距离最小下的时序对齐.尽管采用全局约束,但忽略具体匹配对象的合适性,可能导致文本不同音频序列之间距离较小的情况.因此,基于QbyE方法的关键词检测性能通常比较差.文献[18]针对低数据资源的应用场景提出一种区分性的关键词检测方法,使用直方图矢量来表示孤立词,直方图矢量的获得依赖一个预训练的GMM模型,并在时间帧上做统计平均,忽略关键词的时序信息,因此具有较高的虚警率.

综上所述,针对低数据资源的应用场景,本文借鉴物理空间中的"定位"思想,结合"词是由更小的语言单元,如音节或音素的结构化表达构成,以及任何语言单元具有稳定的统计声学特征"的事实,将语言单元的声学空间视为完整语音空间的子集,多个子空间串构成词的时序,提出一种基于特征空间轨迹信息的语音关键词检测方法.利用大规模无标注的语音样本构建音频特征空间标识子,则标识子可以作为音频特征的位置参照标识.关键词的一次发音视为在音频特征空间中的一次运动,关键词的组成语言单元,如音节或音素,在音频特征空间中的表达序列构成关键词的时序表达,使关键词的空间分布特征和时序特征得到统一描述.由于关键词语音样本仅用于标记该类音频的特征空间分布和时序特征,即使只有少量的样本,也可以进行标记得到基于聚类层次的关键词模板,从而给低数据资源带来泛化能力.本文算法在低数据资源下与基线系统相比取得了更好的检测性能.

2 算法设计

基于上述思路,算法设计首先需要构建一个通用的语音特征空间表达;然后在此特征空间上构建语音关键词的轨迹信息,即空间分布、时序、局部区分信息,以及利用这些信息的关键词检测流程[19].

2.1 构建音频特征空间表达

音频特征空间表达描述一个完整的听觉感知空间,采用聚类算法对语音样本的特征集进行聚类形成音频特征空间的离散标识,这些离散标识子包腔构建音频特征空间,与真实信号空间的差异主要取决于所使用样本对特征空间的覆盖程度,样本越丰富,空间表达越完整.

本文通过EM算法[20]训练GMM模型构建音频特征空间表达.首先从数据集中挑选足够多的语音样本构建音频特征空间表达,挑选的样本集包含数据集中所有说话人且语音内容尽可能丰富,提取梅尔频率倒谱系数(Mel Frequency Cepstum Coefficient,MFCC)特征.然后,利用K-均值(K-Means)算法对特征集进行聚类,计算每个类的均值和方差,作为GMM的初始化参数.最后,通过EM算法求解GMM模型参数,用其高斯分量$\boldsymbol{\Omega}=\{g_k,k=1,2,\cdots,K\}$构建音频特征空间,称$g_k=N(\boldsymbol{m}_k,\boldsymbol{U}_k)$为特征空间标识子,$K$表示音频特征空间标识子的数量.

2.2 构建关键词知识:特征空间轨迹信息

2.2.1 特征$\boldsymbol{f}$的δ邻域标识子

特征$\boldsymbol{f}$表示语音样本的某一帧MFCC特征向量,通过聚类算法获得的音频特征空间标识子$\boldsymbol{\Omega}=\{g_k,k=1,2,\cdots,K\}$具有描述特征$\boldsymbol{f}$在音频特征空间中位置的能力.但是由于不清楚各个标识子之间的位置关系,不利于后续对特征$\boldsymbol{f}$的定位搜索.如果基于某种规则对空间

标识子进行排序,有利于减少计算复杂度,提高搜索速度. 只有特征$\boldsymbol{f}$附近的空间标识子才能为特征$\boldsymbol{f}$提供有意义的位置信息,因此基于特征维度准则构建特征$\boldsymbol{f}$的δ邻域标识子,即对于D维的特征空间,特征$\boldsymbol{f}$在空间中的位置可以根据其与周围(D+1)个标识子的距离确定,可以选择特征$\boldsymbol{f}$的δ邻域大小为(D+1)个邻近标识子. 依据特征$\boldsymbol{f}$与标识子g_k的距离大小进行排序,从而可选择与特征$\boldsymbol{f}$距离最近的(D+1)个标识子作为特征$\boldsymbol{f}$的δ邻域标识子.

(1)计算特征$\boldsymbol{f}$与标识子g_k的均值$\boldsymbol{m}_k$的欧氏距离[21]$d_k(\boldsymbol{f},\boldsymbol{m}_k)$,即

$$d_k(\boldsymbol{f},\boldsymbol{m}_k)=\sqrt{(\boldsymbol{f}-\boldsymbol{m}_k)^{\mathrm{T}}(\boldsymbol{f}-\boldsymbol{m}_k)} \tag{1}$$

其中,$k=1,2,\cdots,K$.

(2)对$d_k(k=1,2,\cdots,K)$按升序进行排序,取前(D+1)个空间标识子作为特征$\boldsymbol{f}$的δ邻域标识子集Φ.

2.2.2　关键词特征空间分布及运动轨迹

基于音频特征空间标识子$\boldsymbol{\Omega}$构建语音样本特征空间分布的过程如下.

(1)提取语音样本的MFCC特征序列$\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_T\}$,构建特征$\boldsymbol{f}_t$的$\delta$邻域标识子集$\Phi_t$.

(2)计算特征$\boldsymbol{f}_t$与标识子$g_k(\boldsymbol{m}_k,\boldsymbol{U}_k)$的位置关联度,位置关联度的定义如下:

$$p_t^k=\begin{cases}\dfrac{1}{\left[(\boldsymbol{f}_t-\boldsymbol{m}_k)^{\mathrm{T}}\boldsymbol{U}_k^{-1}(\boldsymbol{f}_t-\boldsymbol{m}_k)\right]^2}, & g_k\in\Phi_t\\ 0, & g_k\notin\Phi_t\end{cases} \tag{2}$$

其中,$\boldsymbol{f}_t$表示第t帧MFCC特征向量,$\boldsymbol{m}_k$是标识子g_k的均值向量,$\boldsymbol{U}_k$是标识子g_k的对角方差向量,Φ_t是特征$\boldsymbol{f}_t$的δ邻域标识子集.

(3)语音样本的K维特征空间分布可以表示为$\boldsymbol{P}=\left[p^1,p^2,\cdots,p^K\right]^{\mathrm{T}}$,其中,

$$p^k=\frac{\sum_t p_t^k}{\sum_k\sum_t p_t^k} \tag{3}$$

关键词的K维特征空间分布由其语音样本通过上述计算过程获得,词的声学特征统计意义上的稳定性使其空间分布$\boldsymbol{P}$具有区分关键词的能力.

关键词帧级特征序列可视为发音过程在音频特征空间中的运动轨迹,但关键词语音实现的多样性使帧级特征序列难以直接应用于判决. 为了获得关键词音频特征空间运动轨迹的统一描述,方便应用于直接判决或作为模型的输入,本文以音节作为基本单元,通过VAD算法[22]将关键词音频段进行音节划分,获得关键词音节在音频特征空间$\boldsymbol{\Omega}=\{g_k,k=1,2,\cdots,K\}$中的分布序列,该序列保留了关键词音节层面的时序信息.

2.2.3　构建局部区分性信息

局部区分性信息用于区分声学相近、识别易于混淆的词语,如"gōng sī(公司)"和"gōng shì(公式)". 从语义上看,区分性仅来源于"s"和"sh". 它们的特征空间分布信息及音节表达的时序信息可能高度相似,且"s"和"sh"的声学差异可能低于其他相同语言单元"gōng""i"的声学差异之和,从而造成误识. 此时着眼于"s"和"sh"的差异,二者是可以区分的. 另外,音素的发音受上下文等各种因素的影响,其音频信号难以用VAD进行分割. 基于语言单元声学空间的概念,它们的声学区别应反映在其所属的大语言单元(音节)的空间表达中. 因此,可以利用"gōng sī"和"gōng shì"第二个音节特征空间分布的差异构建两者的局部区分性信息,从而提高识别的准确率.

关键词局部区分性信息的构建方法如下:

(1)计算声学混淆音节A和B的特征空间分布$\boldsymbol{P}_A$和$\boldsymbol{P}_B$在每一维上的相对差异得到$\left[p_\Delta^1,p_\Delta^2,\cdots,p_\Delta^K\right]^{\mathrm{T}}$,其中,

$$p_\Delta^k=\frac{\left|p_A^k-p_B^k\right|}{\max\left(p_A^k,p_B^k\right)} \tag{4}$$

p_Δ^k越大,两个音节特征空间分布$\boldsymbol{P}_A$和$\boldsymbol{P}_B$在第k维的差异性越大,则第k个标识子越能作为两者之间的区分性标识.

(2)对$p_\Delta^1,p_\Delta^2,\cdots,p_\Delta^K$从大到小进行排序,取Top-$M$对应的标识子作为两者之间的局部区分性标识子,得到局部区分性标识子集Ψ.

(3)根据局部区分性标识子集Ψ构建掩码向量$\boldsymbol{Q}=\left(q^1,q^2,\cdots,q^K\right)^{\mathrm{T}}$,其中,

$$q^k=\begin{cases}1, & g_k\in\Psi\\ 0, & g_k\notin\Psi\end{cases} \tag{5}$$

任何一个音节都会有声学混淆音节,如果对关键词的每一个音节都构建局部区分性信息,那么局部区分性信息的建设非常庞大. 因此,关键词局部区分性信息的构建和使用是基于应用场景,根据关键词的检出效果和出错情况,有选择性地为关键词构建局部区分信息. 一个关键词添加局部区分信息对其他关键词的检测不产生影响.

2.3　检测流程

本文对连续语音流进行分段识别,对测试样本通过VAD算法进行音节的划分,再根据检测对象构建待检测音段,最终根据测试样本所有待检测音段的识别结果判断该样本是否存在关键词. 测试样本的检测流程图如图1所示.

待检测音段的具体划分策略如图2所示,对测试样本进行音节切分后,选择与预定义关键词的音节数相

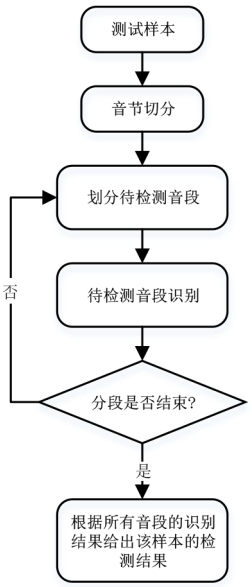

图1 测试样本检测流程图

同的音段作为待检测音段. 例如:测试样本“环保科技有限公司”进行音节切分后有8个音节,与之匹配的关键词“公司”有两个音节,则测试样本最终可以划分7个待检测音段. 由于连续语音的协同发音需要,部分音节的语音划分是很难的,也就是说VAD方法并不能保证将连续音段划分成与对应文字音节数相同的音段. 所以在构建检测音段时,我们还考虑了待检测音段的时长约束:同一检测目标的检测音段时长大致相同.

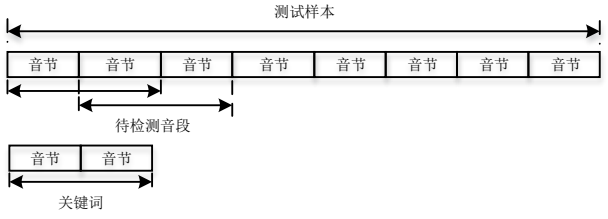

图2 待检测音段的划分策略图

待检测音段与某个关键词的匹配识别过程如图3所示. 待检测音段采用先整体后局部的思路与关键词类进行比较识别,根据比较结果判断待检测音段是否是关键词.

首先,本文基于支持向量机(Support Vector Machines,SVM)[23]判断待检测音段特征空间分布与关键词特征空间分布的相似性,使用关键词和非关键词的特征空间分布矢量集训练SVM分类器. 相比关键词样本,非关键词的样本量要远远大于关键词样本量. 在训练SVM分类器时,一方面,极度不平衡的数据会导致分类器偏向非关键词类,导致分类器不能很好地识别关键词类;另一方面,若只使用与关键词集大小相等的非

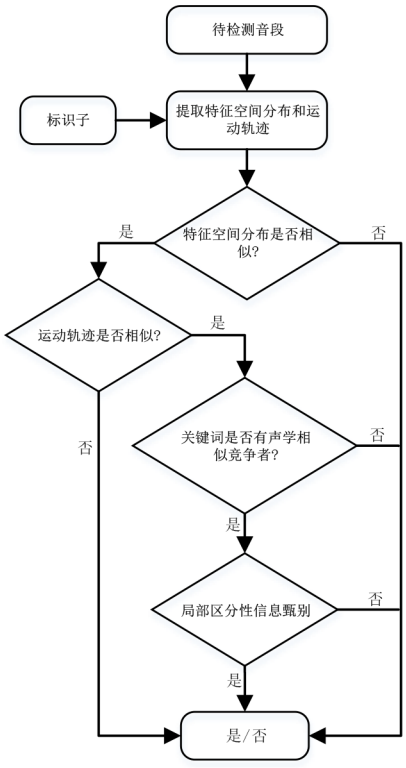

图3 待检测音段与关键词的匹配识别流程图

关键词集进行训练,非关键词样本不够丰富导致分类器不能很好地识别非关键词类. 因此,本文借鉴文献[24]提出的bagging预测器,从非关键词集中随机挑选非关键词子集,其大小与关键词集大小一致,重复采样L次得到L个非关键词子集,每一个非关键词子集与关键词集训练一个SVM二分分类器. 最终,每个关键词都得到L个SVM分类器,对待检测音段识别时进行投票判决. 若所有关键词分类器的判决结果均为负例,则该段的识别结果为非关键词;若分类器的判决结果为正例,则对应的关键词作为待检测音段的候选结果集$W^{(c)}$.

其次,根据待检测音段的特征空间运动轨迹和候选结果集$W^{(c)}$中每个词的特征空间运动轨迹的相似性选出最可能的关键词W_o. 待检测音段与关键词的特征空间运动轨迹的相似性得分定义为

$$\text{score}=\beta\sum_{i=1}^{I}\cos\left(\boldsymbol{P}_i,\boldsymbol{P}_{w_i}\right)+(1-\beta)\sum_{i=1}^{I-1}\sum_{k_1=1}^{K}\sum_{k_2=1}^{K}P_i^{k_1}\times P_{i+1}^{k_2}\times A_{w_{i,i+1}}^{k_1,k_2} \tag{6}$$

其中,$\cos\left(\boldsymbol{P}_i,\boldsymbol{P}_{w_i}\right)$表示计算$\boldsymbol{P}_i$和$\boldsymbol{P}_{w_i}$的余弦相似度;$\boldsymbol{P}_i$表示待检测音段第$i$个音节的特征空间分布;$\boldsymbol{P}_{w_i}$表示关键词$W$第$i$个音节的先验特征空间分布,为样本集中所有样本的特征空间分布的统计平均. $\boldsymbol{A}_{w_{i,i+1}}$是特征空间运动轨迹的另一种表达方式,表示关键词$W$第$i$个音节

到第$i+1$个音节的标识子转移概率矩阵.

$$\boldsymbol{A}_{w_{i,i+1}}=\left[a_w\left(k_1,k_2\right)\right]_{K\cdot K} \tag{7}$$

$$a_w\left(k_1,k_2\right)=\frac{P_{w_i}^{k_1}\cdot P_{w_{i+1}}^{k_2}}{\sum_{k_2}^{K}P_{w_i}^{k_1}\cdot P_{w_{i+1}}^{k_2}} \tag{8}$$

其中,$P_{w_i}^{k_1}$表示$\boldsymbol{P}_{w_i}$第k_1维的值,$P_{w_{i+1}}^{k_2}$表示$\boldsymbol{P}_{w_{i+1}}$第k_2维的值.

最后,当待检测音段与某个关键词W_o的特征空间分布和特征空间运动轨迹都相似时,如果关键词W_o有声学相似竞争者,则利用局部区分性信息进行进一步甄别. 主要分为两种情况:当关键词W_o的声学相似竞争者不在定义的关键词列表内,关键词W_o与其声学相似竞争者第i个音节相似时,对待检测音段第i个音节的特征空间分布$\boldsymbol{P}_i$和关键词W_o第i个音节的特征空间分布$\boldsymbol{P}_{w_{oi}}$利用掩码向量$\boldsymbol{Q}$进行掩码操作后再计算余弦相似度,根据相似度得分判断待检测音段是否是关键词W_o;当关键词W_o的声学相似竞争者在定义的关键词列表内,关键词W_o与其声学相似竞争者第i个音节相似时,考虑使用分类的方式,利用两者进行掩码操作后的特征空间分布数据训练SVM分类器,根据分类结果判断待检测音段是否是关键词W_o.

3　实验数据

为了检测本文方法的有效性,本文分别在普通话数据集和英文数据集上进行实验.

3.1　普通话数据集

本文所使用的普通话数据集来源于AISHELL-1[25]和AISHELL-2[26]. AISHELL-1语料库共包含170 h朗读式的普通话数据,AISHELL-2语料库共包含1 000 h朗读式的普通话数据. 两个数据集中每个样本的采样率为16 kHz. 本文选取8个关键词,即“生产 shēng chǎn”“女子 nǚ zǐ”“公司 gōng sī”“上海 shàng hǎi”“记者 jì zhě”“电视剧 diàn shì jù”“房地产 fáng dì chǎn”“互联网 hù lián wǎng”. 关键词的孤立词数据从数据集中包含关键词的语音样本中截取得到,非关键词的孤立词数据利用音节切分算法从数据集中不包含关键词的语音样本中截取得到,截取的音段包含2~3个音节. 每个关键词选取500个样本作为训练集,100个样本作为验证集,200个样本作为测试集. 关键词检测的测试数据与截取孤立词数据的语音样本不重合,共计3 200条语音样本. 其中,每个关键词的样本量为200,测试数据中包含关键词与不包含关键词的样本比例为1∶1. 每个关键词在不同数据量下的数据从整个数据集中随机抽取,共得到三组具有差异的训练数据,测试数据保持一致,每组实验数据量的统计结果如表1所示.

表1　普通话数据集统计

数据集		关键词样本数	非关键词样本数	总时长/h
孤立词	训练集	4 000	12 000	2.676
	验证集	800	800	0.257
	测试集	1 600	1 600	0.518
关键词检测的测试数据		1 600	1 600	3.485

3.2　英文数据集

英文数据集是为了验证本文所提出的算法具有知识共享性和迁移性. 本文所使用的英文数据集来自谷歌命令词(Google Speech Commands,GSC)[27]. GSC语料库(Version 1)包含30个词语的64 727个语音样本,每个样本包含一个词语,时长为1 s,采样率为16 kHz. 本文选取8个词语作为关键词,即“happy”“house”“bird”“right”“stop”“down”“four”“left”,其他词语作为非关键词. 实验数据的统计结果如表2所示.

表2　GSC数据集统计

数据集	关键词样本数	非关键词样本数	总时长/h
训练集	2 400	7 200	2.70
验证集	800	800	0.45
测试集	1 600	1 600	0.90

每个关键词选取300个样本作为训练集,100个样本作为验证集,训练集中关键词与非关键词样本的比例为1∶3,验证集中关键词与非关键词样本的比例为1∶1,测试集共包含3 200个样本,关键词与非关键词样本的比例为1∶1.

4　实验结果及分析

4.1　评价指标

分类准确率(Acc)定义如下:

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{P+N}\times 100\% \tag{9}$$

其中,TP表示被正确划分为正例的个数,TN表示被正确划分为负例的个数,P表示实际的正例个数,N表示实际的负例个数.

错误拒绝率(False Reject Rate,FRR)和虚警率(False Alarm Rate,FAR)是关键词检测常用的性能指标[10]. 二者定义如下:

$$\mathrm{FRR}=\frac{P-\mathrm{TP}}{P} \tag{10}$$

$$\mathrm{FAR}=\frac{\mathrm{FP}}{\#\mathrm{KW}\times\mathrm{HR}\times C} \tag{11}$$

其中,$P-\mathrm{TP}$表示关键词漏检次数,P表示关键词出现的次数,FP表示关键词虚警次数,#KW表示关键词的个数,HR表示测试样本总时长(单位:h),C是一个常数,本文C=1. FRR和FAR的值越低越好.

错误率(Er)的定义如下:

$$\mathrm{Er}=\frac{\mathrm{FP}}{N}\times 100\% \tag{12}$$

其中,FP表示被错误识别为正例的个数,N表示实际的负例个数.

4.2 实验参数设置

语音分帧长度20 ms,帧移10 ms,利用librosa音频处理库提取12维的梅尔频率倒谱系数(MFCC)特征,即2.2.1节提到的D取12. 构建音频特征空间表达的数据来自AISHELL-1和AISHELL-2数据集,共计24 000条样本,样本与关键词训练集、测试集不重合,音频特征空间标识子数K取4 096. 每个关键词训练L个SVM二分分类器,本文L取5. 使用Python中Scikit-learn模块训练SVM,核函数选择线性核函数. 特征空间运动轨迹的相似性得分计算加权系数β取0.2,构建局部区分性信息时,区分性标识子个数M=256.

本文基于VAD的音节切分算法在连续语音中确定每个音节的浊音起始帧后,根据语音帧之间短时能量和短时过零率的变化向前搜索清音帧的起始点,完成清浊音段的划分. 根据普通话的发音特性,清音在前、浊音在后,将清浊音段进行结合得到音节,从而完成音节的划分. 从AISHELL-1数据集中随机选取100个测试样本,每个测试样本来自不同的说话人,样本总时长大约为6 min,并对样本的音节边界进行人工标注,用于测试音节切分算法的性能. 切分的正确率定义为正确切分的音节边界数与测试所有音节边界数的比值. 其中,算法切分出的音节边界与人工标注的音节边界差距在50 ms以内定义为正确切分. 音节切分的正确率为90.2%,并在此基础上进行实验.

实验中用于对比的基线系统有3个:(1)端到端的CRNN系统,其网络参数与文献[16]一致,包括3层CNN,2层RNN以及2层全连接层,采用CTC(Connectionist Temporal Classification)损失函数;(2)HMM系统,为每个关键词训练一个HMM模型,为其他所有的非关键词训练一个HMM垃圾模型,测试时比较待检测音段在关键词模型和垃圾模型上的得分进行判断;(3)直方图系统[18],使用预训练的GMM模型提取直方图矢量代表孤立词,SVM作为孤立词分类器,对测试样本进行分段识别.

后文中使用FSD表示检测过程只基于SVM判断待检测音段与关键词特征空间分布的相似性,根据SVM的分类结果给出待检测音段的识别结果. 使用FST表示检测过程在FSD的基础上,通过计算特征空间运动轨迹的相似性进行时序判别,但在检测过程中不使用局部区分性信息进行进一步甄别. 使用FST_D表示检测过程在FST的基础上加入局部区分性信息的判别,待检测音段的识别流程如图2所示.

4.3 算法有效性验证

由于本文对测试样本进行分段识别,因此孤立词分类准确率可以反映算法的有效性,分别在每个关键词的训练数据量为10,50,100,300,500时,对比FSD与HMM系统、直方图系统的孤立词分类准确率Acc,如表3所示.

表3 孤立词分类准确率 单位:%

方法	数据量				
	10	50	100	300	500
HMM	50.3	64.9	77.6	86.5	**92.6**
直方图	70.1	78.6	83.9	86.6	87.2
FSD	**73.5**	**82.3**	**88.7**	**91.3**	92.2

根据表3中数据,FSD在低训练数据量下,孤立词Acc优势明显. 在训练数据量为10时,FSD相比于HMM系统,Acc提高了23.2%;相比于直方图系统,Acc提高了3.4%. 随着训练数据量的增加,FSD的孤立词分类准确率始终高于直方图系统,当训练数据量增加到500时,HMM系统的Acc最高,相比于HMM系统,FSD不再具有优势,但性能相当. 说明本文提出的特征空间分布特征具有区分不同词语的能力,且在低数据资源下,FSD相比于基线系统,Acc提升明显.

4.4 训练数据量对关键词检测的影响

为了验证FST_D在低数据资源下的优势,分别在每个关键词的训练数据量为1,10,50,100,300,500时,对比FST_D与基线系统的关键词检测性能. 利用3组训练数据得到的模型对相同的测试数据进行关键词检测得到3组测试结果,3组测试结果存在差异,但差异不大. 表4和表5分别为不同训练数据量下的关键词检测FRR和FAR的平均值,加粗数据为最优结果.

表4 不同训练数据量下关键词检测错误拒绝率

方法	数据量					
	1	10	50	100	300	500
CRNN	—	—	—	—	0.662	**0.105**
HMM	—	0.521	0.375	0.257	0.165	0.106
直方图	0.568	0.356	0.268	0.195	0.152	0.150
FST_D	**0.527**	**0.316**	**0.217**	**0.150**	**0.124**	0.116

表5　不同训练数据量下关键词检测虚警率　　单位:FP/h

方法	数量					
	1	10	50	100	300	500
CRNN	—	—	—	—	**0.66**	**3.14**
HMM	—	21.22	15.95	10.76	7.62	6.17
直方图	28.31	18.24	13.47	9.92	7.68	7.09
FST_D	**20.07**	**12.52**	**7.89**	**5.13**	4.08	3.71

基于表4和表5的数据,在低资源时(100个样本以下),相比HMM和CRNN,FSFD有显著优势,如训练数据量为10时,FST_D相比HMM系统,FRR绝对下降了20.5%,FAR绝对下降了8.7 FP/h. 而CRNN系统在训练数据量低于300时,关键词样本太少导致模型无法收敛,没有检测结果. 在不同数据量下,FST_D的关键词检测性能始终优于直方图系统. 在样本量较充分,训练数据量增加到500时,FST_D的关键词检测性能与CRNN系统相当,FST_D的FRR和FAR略高于CRNN系统,FST_D的FRR略高于HMM系统,FAR相比HMM系统有明显降低.

根据表5,当训练数据从300增加到500时,CRNN系统的FAR变大,为进一步分析,重新选取一组测试数据,在训练数据量分别为300,400,500,600时测试CRNN系统的性能,如表6所示.

表6　不同数据量下CRNN性能

指标	数据量			
	300	400	500	600
FRR	0.677	0.235	0.107	0.103
FAR	0.550	3.570	3.040	2.780

根据表6,随着训练数据量的增加,CRNN系统的FRR呈下降趋势,FAR整体也呈下降趋势.

由于CRNN系统是神经网络模型,需要充足的数据才能使模型收敛到较优的状态. CRNN系统的性能受数据量的影响比较大. 当训练数据增加时,模型的性能提升明显,但低数据资源下,模型无法训练充分,导致性能很差. HMM系统同样需要充足的数据才能使模型训练充分,因此随着训练数据量的增加,FRR下降明显. 同时,由于HMM系统中单一的垃圾模型很难代表所有的非关键词,因此系统的FAR始终比较高. 直方图系统,使用区分性方法,在训练数据量比较少的情况下,与CRNN系统和HMM系统相比具有优势. 但是由于系统使用归一化后的直方图矢量代表孤立词,忽略孤立词的时序信息,因此系统的FAR也比较高,FST_D系统性能始终优于直方图系统. 本文的FST_D方法充分考虑关键词的统计特征以及时序特征,同时SVM分类器的决策依赖支持向量,其分类性能对数据量具有一定的鲁棒性,在数据量较少时,SVM分类器也能取得较优的性能.

总结来说,在低资源时(100个样本以下),FST_D相比HMM和CRNN有显著优势,而在样本量较充分时,与CRNN系统和HMM系统相比,FST_D不再具有优势,但性能大致相当.

4.5　时序信息的作用

本文利用特征空间运动轨迹表达关键词的时序信息,计算特征空间运动轨迹的相似性进行时序判别. 本文选取4组具有时序差异的词语对,即“上海 shàng hǎi”和“海上 hǎi shàng”,“女子 nǚ zǐ”和“子女 zǐ nǚ”,“生产 shēng chǎn”和“产生 chǎn shēng”,“家人 jiā rén”和“人家 rén jiā”. 对“上海 shàng hǎi”“女子 nǚ zǐ”“生产 shēng chǎn”“家人 jiā rén”建立模型,对包含“海上 hǎi shàng”“子女 zǐ nǚ”“产生 chǎn shēng”和“人家 rén jiā”的各100各语音样本进行检测,表7为各种方法的识别错误率. 该实验结果表明,FST利用时序信息大幅降低了错误率,每个词的识别错误率都下降了60%左右,可以说明时序信息的利用对关键词检测任务非常重要,可以在检测过程中降低虚警率.

表7　识别错误率　　单位:%

方法	测试词			
	hǎi shàng	zǐ nǚ	chǎn shēng	rén jiā
HMM	86.0	78.0	77.0	79.0
直方图	84.0	90.0	75.0	86.0
FSD	82.0	88.0	75.0	82.0
FST	**17.0**	**21.0**	**16.0**	**18.0**

4.6　局部区分性信息有效性验证

本文通过构建局部区分性信息来对具有声学混淆性的词语进行甄别,为验证局部区分性信息的有效性,本文选取4组具有声学混淆性的词语对,即“公司 gōng sī”和“gōng shì”,“互联网 hù lián wǎng”和“wù lián wǎng”,“执行 zhí xíng”和“zì xíng”,“资质 zī zhì”和“zhi zhi”. 用“公司 gōng sī”“互联网 hù lián wǎng”“执行 zhí xíng”“资质 zī zhì”的模型对含有其声学混淆词的语音样本进行关键词检测. 每个词语的测试数据

量为200,表8为各种方法识别混淆词的错误率.该结果表明,FST_D可有效地降低声学混淆词识别成关键词.

表8 混淆词识别错误率 单位:%

方法	测试词			
	gōng shi	wù liáng wǎng	zì xíng	zhi zhi
HMM	70.5	96.0	82.0	72.0
直方图	64.5	92.5	76.5	80.0
FST	55.0	90.5	78.5	82.5
FST_D	**4.0**	**20.5**	**21.0**	**18.5**

4.7 音频特征空间表达的可迁移性

本文基于音频特征空间表达构建关键词知识的过程中,基于特征维度准则描述特征在音频特征空间中的位置,是一种相对定位.因此,音频特征空间表达的使用可以不受语种的约束,很好地实现了数据知识的迁移.为验证音频特征空间表达的可迁移性,在GSC数据集上对比分别使用AISHELL和LibriSpeech数据集构建音频特征空间表达时的关键词识别性能.表9为每个关键词的训练数据量为100和300时的分类准确率.

表9 关键词分类准确率 单位:%

方法	数据量	
	100	300
GSC(A)	82.6	83.5
GSC(L)	83.4	83.9

如表9所示,GSC(A)表示GSC数据构建关键词知识所使用的音频特征空间表达是使用AISHELL中文数据集构建的,GSC(L)表示GSC数据构建关键词知识所使用的音频特征空间表达是使用LibriSpeech英文数据集构建的.由表中数据可知,两种情况下,在GSC测试集上的关键词分类准确率相差不大.每个关键词的训练数据量为100时,GSC(A)相比于GSC(L),Acc相差0.81%.说明使用中文数据集构建的音频特征空间表达具有区分不同类别英文数据的能力,音频特征空间表达的使用可以不受语种的约束,具有可迁移性.

5 结束语

本文基于特征空间轨迹信息的语音关键词检测方法,利用了关键词的音频统计特征、时序特征以及与相近声学表征语言单元的局部区分信息.多个实验表明,该方法具有良好的适应不同资源的性能表现,特别是对于低资源场景.基于物理定位原理的统计特征比文献[18]的直方图表达更准确,而时序信息可区分成分相同、时序不同的词,即具有判别AB与BA的能力.利用不同音频类之间的局部区分性信息,对辨别声学相近、易于混淆的词语有利.音频特征空间的建立由于不依赖具体关键词的样本数量,本质上与语言关联性小,特征空间表达具有语言迁移特性.关键词样本用来标记该类音频运动轨迹信息,即使只有少量的样本,也可得到聚类水平模板的表达,从而给低数据资源带来泛化能力.基于中文数据AISHELL和英文数据GSC的实验结果表明:在低资源时(100个样本以下),相比HMM,CRNN有显著优势;而在样本量较充分(300及以上)时,与CRNN有大致相当的性能.

若从关键词概念层面考虑,任何多音节关键词均可表达为音节序列(实际应用中,单个汉字极少成为关键词),因此本文方法具有普适性.对于非中文语言,比如英语,将“音节”理解为“最小清浊音段”,本文方法的描述就适用于英语等拼音语言.

参考文献

[1] SANGEETHA J, JOTHILAKSHMI S. A novel spoken document retrieval system using auto associative neural network based keyword spotting[C]//2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO). Piscataway: IEEE, 2015: 1-6.

[2] 刘俊华.面向多语种海量数据的语音关键词检索方法研究与系统实现[D].合肥:中国科学技术大学, 2019.
LIU J H. Research and System Implementation of Speech Keyword Retrieval Method for Multilingual Massive Data [D]. Hefei: University of Science and Technology of China, 2019. (in Chinese)

[3] KAVYA H P, KARJIGI V. Sensitive keyword spotting for crime analysis[C]//2014 IEEE National Conference on Communication, Signal Processing and Networking (NCCSN). Piscataway: IEEE, 2015: 1-6.

[4] CHEN G G, PARADA C, HEIGOLD G. Small-footprint keyword spotting using deep neural networks[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2014: 4087-4091.

[5] MICHAELY A H, ZHANG X D, SIMKO G, et al. Keyword spotting for Google assistant using contextual speech recognition[C]//2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Piscataway: IEEE, 2018: 272-278.

[6] WEINTRAUB M. LVCSR log-likelihood ratio scoring for keyword spotting[C]//1995 International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 1995: 297-300.

[7] ROSENBERG A, AUDHKHASI K, SETHY A, et al. End-to-end speech recognition and keyword search on low-resource languages[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2017: 5280-5284.

[8] 唐海桃, 薛嘉宾, 韩纪庆. 一种多尺度前向注意力模型的语音识别方法[J]. 电子学报, 2020, 48(7): 1255-1260.
TANG H T, XUE J B, HAN J Q. A method of multi-scale forward attention model for speech recognition[J]. Acta Electronica Sinica, 2020, 48(7): 1255-1260. (in Chinese)

[9] ROSE R C, PAUL D B. A hidden Markov model based keyword recognition system[C]//International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE, 1990: 129-132.

[10] ZHANG S L, SHUANG Z W, SHI Q, et al. Improved mandarin keyword spotting using confusion garbage model[C]//2010 20th International Conference on Pattern Recognition. Piscataway: IEEE, 2010: 3700-3703.

[11] CHEN Q Y, ZHANG W B, XU X M, et al. Improved keyword spotting based on keyword/garbage models[C]//2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Piscataway: IEEE, 2017: 1-4.

[12] SIGTIA S, HAYNES R, RICHARDS H, et al. Efficient voice trigger detection for low resource hardware[C]//Interspeech 2018. Baixas: ISCA, 2018: 2092-2096.

[13] SHRIVASTAVA A, KUNDU A, DHIR C, et al. Optimize what matters: Training DNN-hmm keyword spotting model using end metric[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2021: 4000-4004.

[14] ARIK S O, KLIEGL M, CHILD R, et al. Convolutional recurrent neural networks for small-footprint keyword spotting[EB/OL]. (2017-03-15)[2022-03]. https://arxiv.org/abs/1703.05390.

[15] WANG Y Y, LONG Y H. Keyword spotting based on CTC and RNN for mandarin Chinese speech[C]//2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP). Piscataway: IEEE, 2019: 374-378.

[16] YAN H K, HE Q H, XIE W. Crnn-ctc based mandarin keywords spotting[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2020: 7489-7493.

[17] MADHAVI M C, PATIL H A. Vocal Tract Length Normalization using a Gaussian mixture model framework for query-by-example spoken term detection[J]. Computer Speech & Language, 2019, 58: 175-202.

[18] BENISTY H, KATZ I, CRAMMER K, et al. Discriminative Keyword Spotting for limited-data applications[J]. Speech Communication, 2018, 99: 1-11.

[19] 贺前华, 田颖慧, 兰小添, 等. 一种基于运动轨迹和区分性信息的语音关键词检测方法: CN114373453A[P]. 2022-04-19.
HE Q H, TIAN Y H, LAN X T, et al. Voice keyword detection method based on motion trail and distinguishing information: CN114373453A[P]. 2022-04-19. (in Chinese).

[20] KHELIFA M O M, ELHADJ Y M, ABDELLAH Y, et al. Constructing accurate and robust HMM/GMM models for an Arabic speech recognition system[J]. International Journal of Speech Technology, 2017, 20(4): 937-949.

[21] 唐宇政. 基于欧式距离的判别分析: 鸢尾花分类问题探究[J]. 现代商贸工业, 2019, 40(9): 183-185.
TANG Y Z. Discriminant analysis based on euclidean distance-Research on iris classification[J]. Modern Business Trade Industry, 2019, 40(9): 183-185. (in Chinese)

[22] 贺前华, 苏健彬, 严海康, 等. 一种基于语谱图时间差分的语音音节数估计方法: CN111063371A[P]. 2023-04-21.
HE Q H, SU J B, YAN H K, et al. Speech Syllable Number Estimation Method Based on Spectrogram Time Difference: CN111063371A[P]. 2023-04-21. (in Chinese).

[23] NOBLE W S. What is a support vector machine?[J]. Nature Biotechnology, 2006, 24(12): 1565-1567.

[24] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[25] BU H, DU J Y, NA X Y, et al. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline[C]//2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). Piscataway: IEEE, 2018: 1-5.

[26] DU J Y, NA X Y, LIU X C, et al. AISHELL-2: Trans-

forming mandarin ASR research into industrial scale [EB/OL]. (2018-08-31)[2022-03]. https://arxiv.org/abs/1808.10583.

[27] WARDEN P. Speech commands: A dataset for limited-vocabulary speech recognition[EB/OL]. (2018-04-09)[2022-03]. https://arxiv.org/abs/1804.03209.

作者简介

田颖慧 女,1997年生,河南驻马店人. 华南理工大学硕士研究生. 主要研究方向为语音信号处理、语音关键词检.
E-mail: 13174416712@163.com

贺前华(通讯作者) 男,1965年生,湖南邵东人. 博士. 华南理工大学教授、博士生导师. 主要研究方向为智能音频信号处理、语音识别和说话人识别.
E-mail: eeqhhe@scut.edu.cn

郑若伟 男,1998年生,广东汕头人. 华南理工大学硕士研究生. 主要研究方向为语音关键词检测、语音识别.
E-mail: ruoweizheng@foxmail.com

危 卓 女,1997年生,湖南岳阳人. 华南理工大学硕士研究生. 主要研究方向为语音信号处理、说话人识别.
E-mail: 201921011738@mail.scut.edu.cn

李艳雄 男,1980年生,湖南嘉禾人. 博士. 华南理工大学副教授、博士生导师. 主要研究方向为语音及音频信号处理、机器学习.
E-mail: eeyxli@scut.edu.cn