ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 2,Mar. 2017,pp. 346－353
DOI:10.16337/j.1004-9037.2017.02.016

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于得分归一化和系统融合的语音关键词检测方法

李　鹏　屈　丹

(解放军信息工程大学信息系统工程学院，郑州，450001)

摘　要：为了有效利用不同关键词检测系统的互补性，解决不同系统检测结果置信度得分不在同一范围的问题，提出了一种基于得分规整和系统融合的语音关键词检测方法。首先，为了克服连续语音识别系统中因剪枝错误而引起的关键词丢失问题，应用了关键词相关的软 Beam 宽度剪枝策略裁剪词图；其次，在系统融合前采用得分归一化方法，使得不同系统关键词检测结果置信度得分在同一范围；最后，通过系统融合处理将不同系统的关键词输出进行整合，得到最终的关键词检测结果。实验结果表明，经过得分归一化处理后，关键词检测性能的实际查询词权重代价(Actual term-weighted value，ATWV)平均相对提升 30%；系统融合后关键词的检测性能，相比于得分归一化处理后的最佳单一系统，得到了 10% 的提升。

关键词：关键词检测；得分归一化；系统融合；软 Beam 剪枝

中图分类号：TP391　　**文献标志码**：A

Keyword Spotting based on Score Normalization and System Combination

Li Peng，Qu Dan

(School of Information Systems Engineering，PLA Information Engineering University，Zhengzhou，450001，China)

Abstract：To effectively use the complementarity of different keyword spotting systems and solve the problem that the confidence scores from several different subsystems is not in the same range，a keyword spotting system based on score normalization and system combination is proposed. Firstly，to avoid keyword missing due to pruning errors in a large vocabulary recognition system，the keyword soft Beam pruning method is presented. Secondly，score normalization is needed to transform these confidence scores into a common domain，prior to combining them. Finally，after score normalization，the outputs are combined from different subsystems. Results show that score normalization methodology improves keyword search performance by 30% in average. Experiment also show that combining the outputs of diverse systems，system perform is 10% better than the best normalized KWS system.

Key words：keyword spotting；score normalization；system combination；soft Beam pruning

引　　言

随着科技的发展，音频数据量急剧增长。特别是广播节目、语音文档以及会议语音记录等语音信息以互联网为载体，传播广泛，急需一种有效的方法来检索这些语音信息。语音关键词检测(Keyword

基金项目：国家自然科学基金(61175017)资助项目。
收稿日期：2014-06-19；**修订日期**：2014-11-18

spotting,KWS)[1]是指在大量语音资料中快速检索并返回关键词精确位置信息的技术。在 KWS 系统中,常用的系统融合方法是在两个子系统中对所有查询词分别进行检索,并将不同子系统的检测结果集合进行合并。具体而言,融合分数的方法可以采用简单的算术平均值[2],也可以对基于音节和词的检索系统的结果集进行线性的分数融合[3]。但已有方法只是两个子系统的建模单元不同,并没有充分利用多系统的互补性,关键词检测系统提升性能有限。

因此,充分利用不同系统的互补性来提高 KWS 系统的性能成为目前研究的热点之一。孟猛等提出的多系统融合语音 KWS 方法[4]是通过应用不同的语音特征及声学建模方法来获得互补系统,该方法在一定程度上提高了 KWS 系统的性能;Lidia Mangu 等[5]研究了连续语音识别系统的多样性对合并 KWS 结果性能的影响;Arindam 等[6]在语音数据噪声大、信噪比低的情况下,应用对噪声具有鲁棒性的特征提取方法,合并多样性和互补性的连续语音识别系统的关键词检测结果,显著地提升了 KWS 性能。但互补性多系统融合语音 KWS 方法会面临两个新的问题,一是不同子系统的 KWS 结果得分常不在同一范围,为了进行有效的后处理,需要进行得分归一化处理;二是现有的分数融合方法相对单一。此外,由于 KWS 系统的性能很大程度上依赖于连续语音识别系统的准确性,因此常使用词图(Lattice)等多候选识别结果建立索引进行关键词的检测。连续语音识别系统在生成 Lattice 的过程中要进行剪枝操作,为了防止剪枝引起的关键词丢失,提高关键词的检出率,Zhang 等提出了软 Beam 剪枝宽度调整策略[7]。在软 Beam 剪枝操作中,如果状态节点属于关键词列表,剪枝宽度会更宽。这样如果关键词存在于语音中,即使得分较低也不会因为剪枝而丢失。

为了进一步提高关键词检测系统的检出率同时尽可能降低虚警率,本文通过应用多样性和互补性的连续语音识别组成部分(特征提取、声学模型和解码方式等)得到不同的连续语音识别中间结果 Lattice,在解码过程中首先进行关键词相关的软 Beam 剪枝;然后对软 Beam 剪枝后的词图分别建立索引,各自独立进行检索;检索结果进行得分归一化处理后,将不同词图得到的关键词检测结果合并到一个列表中并分配相应的得分;最后经过硬判决来决定每个关键词检测结果。为了验证算法的性能,本文进行了微软语料库上的关键词检测实验,比较了不同的归一化方法以及系统合并方法对实际查询词权重代价(Actual term-weighted value, ATWV)的影响。

1　关键词检测系统

KWS 系统的结构如图 1 所示,语音数据经过连续语音识别系统产生词级 Lattice,在连续语音解码过程中应用关键词相关的软 Beam 宽度剪枝;然后将软 Beam 宽度剪枝处理后的 Lattice 转换为基于加权有限状态转换器(Weighted finite state transducer, WFST)的索引结构进行检索;检索结果进行得分归一化处理后,将不同 Lattice 得到的关键词检测结果经过时间对齐,分数融合操作,得到最终系统合并后的 KWS 结果。

1.1　Lattice 生成

系统融合提升关键词检测性能的关键是多套语音识别系统之间具有良好的互补性,即各识别系统性能相当,同时又存在一定的差异性。本文通过构建具有差异性的声学模型来获得互补的识别系统,进而通过融合手段来提高关键词检测系统的性能。搭建了 4 种不同的连续语音识别系统生成 Lattice。

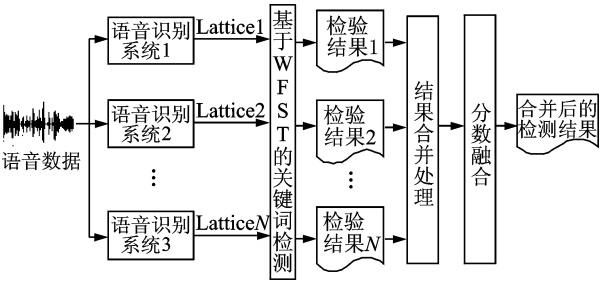

图 1　关键词检测系统

Fig. 1　Keyword spotting system

(1)高斯混合模型(Gaussian mixture model, GMM),采用传统隐马尔可夫模型(Hidden Markov model, HMM)-GMM 框架的连续语音识别系统;(2)LDA+SAT+MLLT,在 HMM-GMM 框架下对

特征进行了线性判别分析(Linear discriminative analysis, LDA),利用最大似然线性变换(Maximum linear likelihood transformation, MLLT)进行说话人自适应训练(Speaker adapted training, SAT)来获得声学模型;(3)基于子空间高斯混合模型(Subspace Gaussian mixture model, SGMM),应用了子空间高斯混合的声学模型,与传统声学模型不同,在 SGMM 中所有的状态共享相同 GMM 结构(相同的高斯混元数和协方差矩阵),并应用了基于最大互信息准则(Maximum mutual information, MMI)区分性训练;(4)DNN,建立一个深层神经网络(Deep neural network,DNN)声学模型,相继进行了预训练,帧级的互熵训练和状态级的最小贝叶斯风险训练。传统的 GMM 模型,应用期望最大化 (Expectation maximization, EM)算法在最大似然度准则(Maximum likelihood criteria,MLC)下,模拟每一种音素分类的分布,但它没有考虑如何更好地区分各类别。SGMM 声学模型则在 GMM 模型的基础上,所有的状态共享相同 GMM 结构(相同的高斯混元数和协方差矩阵),并应用最大互信息区分性准则来优化模型。与 SGMM 与 GMM 模型不同,DNN 模型则直接对各音素类别的后验概率,以最大限度区分各类音素类别为目的进行训练。因此,DNN 模型与 SGMM,GMM 模型从原理上就存在差异性,并且连续语音识别性能相差不大,可以作为互补模型。

1.2 基于 WFST 的索引和检索

实验中应用基于 WFST 的关键词索引和检索方法,语音信号经过大词汇量连续语音识别系统,及软 Beam 剪枝操作后,得到相应的词级 Lattice,并作为关键词检测系统的输入。音节级的 Lattice 由词级 Lattice 转化,用于建立音节索引来检测集外词。

文献[8]详细描述了 Lattice 转化为基于 WFST 索引的算法。首先,将每句话的词级 Lattice 转化为自动机,并提取其中时间信息(各节点到初始节点的时间);然后,对 Lattice 转化而成自动机进行预处理,应用权重推进算法[9]将权重转化为后验概率,经过预处理后,通过因子选择与优化两步构建基于时间的因子转换器;最后将每个句子的因子转换器联合成一个总的因子转换器,并运用加权 ε 移除、确定化或最小化等算法优化得到索引。用 Q 表示一个有限状态集,在基于 WFST 的索引中,Lattice 中的每一条弧由一个 5 元组 (p,i,o,ω,q) 表示。其中 $p \in Q$ 为起始状态,$q \in Q$ 代表终止状态,i 代表输入符号,o 为输出符号,ω 表示 i 的后验概率。

基于 WFST 的检索操作中,每一个查询项用加权接收机来表示,将查询项自动机与已建立的索引作合成运算[10],并进行 ε 移除、确定化或最小化等优化操作得到表示检索结果的自动机。每个查询词被分为集内词和集外词,集内词在词级索引中检索,集外词在音节索引中检索。最后将检索结果统一到一个列表中。检索结果用一个 4 元组表示为 (i_d,s,d,p),其中 i_d 为语音文件序号,s 为检索结果的起始时间,d 为检索结果的持续时间,p 为检索结果的后验概率估计值。

2 得分归一化

系统融合中的关键步骤是对关键词检测结果的置信度得分进行分数融合,然而不同系统的关键词检测结果的置信度得分常不在同一范围,不具有可比性。得分归一化是分数融合的预处理步骤,使得各个系统的检测结果置信度得分具有可比性,从而达到最优的检测性能。关键词检测结果的后验概率、候选检索结果个数、候选检索结果的持续时间、所有候选检索结果持续时间的平均值以及查询项包含字(音节)的个数,都会在一定程度上影响关键词检测的 ATWV 得分。基于以上考虑,本文研究了不同得分归一化方法对关键词检测性能的影响。假设 N_q 为查询项 q 的候选检索结果数,$s_{q,i}$ 为查询项 q 的第 i 个候选检索结果的得分。

(1)查询项长度归一化方法(Query Length, QL)。为了进一步减少误警,应用了基于候选检索结果持续时间的查询项长度归一化方法。由于某一个候选检索结果的持续时间相对较长,正确的可能性越高,因此定义 QL 方法为

$$s_{q,i} = s_{q,i}^{\frac{1}{\text{dur}(q)}} \tag{1}$$

式中:$s_{q,i}$为查询项 q 的第 i 个候选检索结果的归一化得分,$\Delta\text{avg}(q)$为查询项 q 的所有候选检索结果的时间平均值。

(2)累加归一化方法(Sum to one, STO)计算公式为

$$\bar{s}_{q,i}=\frac{s_{q,i}}{\sum_{j=1}^{N_q}s_{q,j}} \tag{2}$$

对于一个给定的查询项,其所有的候选检索结果归一化处理后的分数累加为 1。对于特殊情况,当查询项的候选检索结果只有一个时,归一化分数为 1。对于候选检索结果比较少的查询项,分母会较小,但其归一化得分相对较大,很有可能高于关键词的判决门限,因此漏警率会降低。

(3)基于回归的归一化方法(Pace regression),简称 Pace 回归算法。应用了一种基于机器学习的得分归一化方法,通过融合候选检索结果的以下 5 个特征:后验概率、候选检索结果个数、候选检索结果的持续时间,所有候选检索结果持续时间的平均值和查询项包含字(音节)的个数。采用 Pace 回归算法,学习得分归一化函数来计算归一化得分。

3　系统融合

在不同 KWS 系统的结果合并中,时间对齐与分数融合是要解决的重要问题。首先,如何对不同检索结果集合中的候选位置进行时间对齐,不同系统输出相同检索结果候选的起始和结束时间具有差异性,因此需要将其统一到最终检测结果;其次,如何对不同子系统具有同一检测结果的分数进行融合,分数融合方法对 KWS 性能具有直接影响,因此选取有效的分数融合方法,才能获得更好的 KWS 性能。

3.1　时间对齐

对于某一个查询项,不同子系统产生相同的候选结果,但是不同系统输出相同候选检索结果的起始和结束时间具有差异性,系统融合是将它们合并为一个检测结果。如果一个子系统检测结果中查询词出现在句首,另一个子系统检测结果中的查询词出现在句尾,则两者显然应该作为不同的候选结果进行处理。时间对齐是结果合并的预处理步骤,不同子系统产生相同的候选结果,合并为一个检测结果的条件为:(1)存在于同一条语句,(2)时间上有交叠。合并后检测结果的起始时间和终止时间为合并前得分最高的候选检测结果的起止时间。没有与其他系统交叠的候选检测结果直接拷贝至最后合并的列表。

3.2　分数融合

分数融合步骤综合利用各个子系统的置信度得分信息,使得合并后 KWS 结果的置信度得分更加准确,本文研究了不同分数融合方法对 KWS 性能的影响。其中,h_i 表示在 n 个检索系统中,第 i 个子系统对于查询项 H 能够进行合并的检索结果,$\bar{s}(h_i)$为 h_i 经过归一化处理后的得分,$\hat{s}_i$ 为融合后的得分。

(1)CombSUM 方法[11],将不同系统中可以进行合并的检索结果的得分进行求和,计算公式为

$$\hat{s}_i=\sum_{i=1}^{n}\bar{s}(h_i) \tag{3}$$

(2)CombMNZ 方法[11],计算公式为

$$\hat{s}_i=m_{\mathrm{H}}\times\sum_{i=1}^{n}\bar{s}(h_i) \tag{4}$$

其中 m_{H} 为可以进行合并的检索结果得分不为 0 的数目。

(3)线性合并方法 Linear combination(CombLC),是对 CombSUM 方法的扩展,对合并的系统 i 分配一个权重 w_i,计算公式为

$$\hat{s}_i=\sum_{i=1}^{n}w_i\cdot\bar{s}(h_i) \tag{5}$$

其中的关键问题是权重 w_i 的确定。线性回归[12]是目前应用比较广泛的解决权重分配的方法,也可以

根据系统的检索性能分配权重。

(4)WCombMNZ 方法。其为本文中提出的新方法,最大查询项权重代价(Maximum term-weighted value, MTWV)反映了最佳关键词检测门限下的性能指标,用了 MTWV 指标当作权重的 CombMNZ 合并方法,更加合理地分配了子系统的权重,能够获得最佳的关键词检测性能。计算公式为

$$s_i = m_{\mathrm{H}} \times \sum_{i=1}^{n} w_i^{\mathrm{MTWV}} \cdot \bar{s}(h_i) \tag{6}$$

其中

$$w_i^{\mathrm{MTWV}} = \frac{\mathrm{MTWV}_i}{\sum_{j=1}^{n} \mathrm{MTWV}_j} \tag{7}$$

用 MTWV_i 表示第 i 个子系统的 MTWV 得分。

4 实验结果及分析

4.1 实验配置

实验采用微软汉语语料库,其中训练集由 100 个年龄在 18～40 岁间的男性录音(大部分在 25 岁以下)组成,每人大约 200 句,共 19 688 句话,454 315 个音节。测试集包含 25 个不同的男性录音,每人 20 句,共 500 句话。依据测试集的词频,选取 50 个关键词查询项,词频均匀分布。关键词的长度由汉字个数决定,其中一字词至五字词各 10 个。

利用 Kaldi[13] 工具包搭建 GMM, LDA＋SAT＋MLLT、SGMM 和 DNN 连续语音识别系统:(1) GMM 系统,应用 13 维 MFCC 特征及其 1,2 阶差分,特征矢量共 39 维,帧长和帧移分别为 25 ms 和 10 ms。采用倒谱均值方差归一化(Cepstrum mean and variance normalization, CMVN)方法对每一个说话人语音数据的特征矢量进行处理。采用上下文相关三音子(triphone)为声学建模单元,聚类后得到 1 935个不同的绑定状态(Tied states)。(2) LDA＋SAT＋MLLT 系统,在 GMM 系统的基础上对特征进行 LDA,利用 MLLT 进行 SAT 来获得声学模型。(3)SGMM 系统,为在 GMM 系统的基础上应用 Kaldi 工具箱训练基于子状态的扩展 SGMM 声学模型。(4)DNN 系统,DNN 声学模型建立过程中,神经网络设置两个隐含层,DNN 的输出有 1 935 个节点。网络的输入为 9 帧(当前帧的前后各 4 帧信号),每帧为 40 维的特征矢量,并应用线性区分性分析、MLLT 和特征域最大似然线性变换(Feature-space maximum likelihood linear regression, FMLLR),将 40×9＝360 维的特征矢量变为 250 维。利用 Kaldi 工具包相继进行了预训练,帧级的互熵训练和状态级的最小贝叶斯风险训练实现了 DNN 声学建模。测试阶段,采用 Kaldi 基于 WFST 的解码器构建静态解码网络对测试集进行解码识别,并分别生成词图。本文实验均采用无调音节进行解码识别,并生成词级词图,最后经过基于 WFST 的语音查询项检索系统。

4.2 评价标准

应用查询项权重代价(Term-weighted value, TWV)指标来对 KWS 系统性能进行评价,并根据候选检测结果中正确命中,漏警和虚警的数目来进行计算。首先针对每个查询项计算虚警率和漏警率,对两类错误进行有权重的累加,然后在所有的查询词上进行平均。TWV 定义为

$$\mathrm{TWV}(\theta) = 1 - \frac{1}{N} \sum_{term=1}^{N} \left(P_{\mathrm{miss}}(term, \theta) + \beta P_{\mathrm{fa}}(term, \theta) \right) \tag{8}$$

$$P_{\mathrm{miss}}(term) = 1 - \frac{N_{\mathrm{corr}}(term)}{N_{\mathrm{true}}(term)} \tag{9}$$

$$P_{\mathrm{fa}}(term) = \frac{N_{\mathrm{spur}}(term)}{T - N_{\mathrm{true}}(term)} \tag{10}$$

式中:θ 为置信度门限;N 为关键词查询项的总数;N_{corr} 和 N_{spur} 分别为每一个查询项中正确识别和错误

识别的总数;N_{true}为标注中关键词出现的次数;T 为关键词测试语料的时长(以 s 为单位);β 为常量,通常设置为 999.9。TWV 指标是置信度门限 θ 的函数,在不同门限 θ 下,将得到不同的虚警率 $P_{\text{fa}}(term,\theta)$和漏警率 $P_{\text{miss}}(term,\theta)$,TWV 数值也不同。实际门限值下的 TWV 值称为 ATWV,本文应用 ATWV 指标作为关键词检测系统的评价指标。

4.3　单系统连续语音识别性能

表 1 给出了不同连续语音识别系统的识别结果,采用词错误率(Word error rate,WER)和词图密度(Lattice density)来评价其性能。其中词图密度为词图中包含的弧数目与正确标注中词个数的比值。表 1 中 SGMM 的系统性能最优,其次为 DNN,再次为 LDA+SAT+MLLR,HMM-GMM 系统性能最低,WER 为 20.45%。基于 SGMM 的连续语音识别系统中,声学模型状态相关参数较少,它可以利用集外数据对参数子空间进行估计,声学建模过程中参用了区分性训练与说话人自适应训练,因此 WER 相对较低。在 Lattice 的生成过程中采用相同的剪枝门限,由于不同识别系统的声学建模方式不同,产生 Lattice 的词图密度具有一定的差异性。

表 1　连续语音识别系统的词错误率和词图密度

Tab. 1　Performance of four ASR systems measured in terms of WER and lattice density

系统	WER/%	Lattice density
GMM	20.45	1458
LDA+SAT+MLLT	16.31	528
SGMM	13.91	505
DNN	14.49	1055

4.4　软 Beam 宽度剪枝性能分析

语音识别器在不同的 Beam 剪枝宽度下,会产生不同尺寸的 Lattice,识别结果也不同。本文采用 Lattice 错误率(Lattice error rate, LER)[14] 来评价 Lattice 的性能,LER 也称为 ORACLE 错误率,指 Lattice 中和正确句子最相似的识别候选的错误率,它反映了 Lattice 多候选识别结果的错误率下界。

关键词检测的性能本质上由连续语音识别率决定。在基于 WFST 的关键词检索中 LER 直接影响关键词检索性能。LER 越小,Lattice 中包含正确识别结果候选的可能性越大,关键词的检索系统的漏警率越低。以 GMM 连续语音识别系统为例,图 2 给出了随着剪枝门限的不同,传统剪枝方法与软 Beam 剪枝方法的 LER 的变化曲线。由图 2 可以看出软 Beam 剪枝下的 LER 更低,而传统剪枝方法的 LER 会随着剪枝宽度的变大趋于不变。原因在于软 Beam 剪枝操作中,如果状态节点属于关键词列表,剪枝宽度会更宽。这样如果关键词存在于语音中,即使它的得分较低也不会因为剪枝而丢失。因此 Lattice 中存在正确识别结果的可能性更大,LER 相对较低。随着剪枝门限增大,Lattice 尺寸会急剧变大,索引与检索的代价增大,速度也会明显降低。实验中采用软 Beam 剪枝宽度为 300,LER 相对较低,Lattice 尺寸较小。

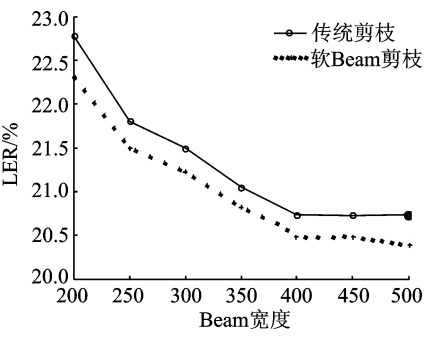

图 2　不同剪枝门限下的 LER

Fig. 2　LER under different pruning thresholds

4.5　得分归一化方法性能分析

表 2 比较了不同得分归一化方法下关键词检测的性能。在各个系统中,分别应用了 QL,Pace 方法和 STO 方法对关键词检索基线系统进行得分归一化处理,其中基线系统为未经过得分归一化处理的 ATWV 得分。

表 2　不同得分归一化方法的 ATWV 得分

Tab. 2　ATWV results of different score normalization methodologies

系统	基线系统	QL	Pace	STO	STO 相对提升/%
GMM	0.311	0.339	0.434	0.470	51.1
LDA+SAT+MLLT	0.405	0.430	0.443	0.538	32.8
SGMM	0.407	0.435	0.461	0.524	28.7
DNN	0.455	0.495	0.520	0.562	23.5

表 2 中最好的关键词检测性能来自于 DNN 系统的 STO 归一化方法。STO 归一化方法相比基线系统平均相对提高 30%；而 QL 方法、Pace 方法分别平均相对提高 9%和 19%。经过得分归一化处理后，关键词检测性能指标 ATWV 都有明显提升。其中 STO 归一化方法性能最佳，原因在于关键词检测中，查询项的候选结果个数差异性较大，置信度分数常不在同一范围。选用全局门限来决定关键词候选结果的正确性，容易产生漏警。对于候选检索结果较少的查询项，STO 方法分母会较小，但是它们的归一化得分相对较大，很有可能高于关键词的判决门限，因此漏警率会降低。QL 方法只考虑了查询项持续时间对关键词检测性能的影响，得分归一化处理后，系统性能提升有限。

不同的系统经过得分归一化后提升的程度不同，如 LDA＋SAT＋MLLT 系统的连续语音词错误率高于 SGMM 系统，但是 STO 归一化处理后，LDA＋SAT＋MLLT 系统的关键词检测性能更佳。原因在于 WER 是在 1-best 上计算，而查询项是在整个 Lattice 上进行检索。

4.6 系统合并方法性能分析

4.2 节提出的不同分数融合方法下的关键词检测性能比较结果如表 3 所示。与归一化方法结合后不同的系统合并方法如下：(1)CombSUM，CombLC 和 CombMNZ 为原始得分直接进行合并；(2)STO-CombLC，STO 归一化后的分数进行 CombLC 合并，权重由线性回归模型确定；(3)STO-CombMNZ，STO 归一化后的分数进行 CombMNZ 方法合并；(4) CombMNZ-STO，原始得分经过 CombMNZ 方法合并后，再进行 STO 归一化处理；(5)WCombMNZ-STO，原始得分经过 WCombMNZ 合并后，再进行 STO 归一化处理；(6)STO-WCombMNZ-STO，STO 归一化处理后的得分经 WCombMNZ 合并后，再经过 STO 归一化处理。

表 3 不同系统合并方法的 ATWV 得分

Tab. 3 ATWV results of different system combination methodologies

系统合并方法	ATWV
CombSUM	0.412
CombLC	0.531
CombMNZ	0.573
STO- CombLC	0.546
STO- CombMNZ	0.602
CombMNZ-STO	0.605
WCombMNZ-STO	0.617
STO- WCombMNZ-STO	0.621

不同系统合并方法，相比单一系统关键词检测性能都具有一定程度的提升。系统合并后再应用得分归一化方法，关键词性能提升更明显。其中 STO-WCombMNZ-STO 合并方法相比传统的合并方法(CombLC，CombSUM，CombMNZ)性能更好。相比最佳的得分归一化方法(DNN 系统的 STO 归一化方法)，相对提高了 10%。WCombMNZ 合并方法的权重，用单系统的 MTWV 得分来计算，充分利用了不同系统的互补性，最大限度提高了合并系统的关键词检测性能。

4.7 查询项长度对关键词检测性能的影响

实验分析了 STO 归一化方法和 STO-WCombMNZ-STO 合并方法，在查询项长度影响下的关键词检测性能。在表 4 中基线系统为 DNN 系统未经过得分归一化处理的 ATWV 得分。表 4 分别给出了 STO 归一化方法相对基线系统的提升比率，以及 STO-WCombMNZ-STO 合并方法相对于 STO 归一化方法的提升比率。

表 4 查询项长度对关键词检测性能 ATWV 的影响

Tab. 4 Influence of ATW on query length

长度	基线系统	STO	STO 提升/%	合并方法	合并提升/%
2	0.073	0.094	28.8	0.102	8.5
3	0.661	0.750	23.5	0.835	11.3
4	0.635	0.706	11.2	0.866	22.6
5	0.536	0.576	7.5	0.683	18.6

从表 4 可以得出如下结论：(1)STO 归一化方法对于不同查询项长度的关键词检测性能都有所提升，但对于较短的查询项提升性能更佳；(2)STO-WCombMNZ-STO 合并方法对于不同查询项长度的关键词检测性能都有所提升，但对于较长的查询项提升性能更佳；(3)对于 4 字词查询项系统合并性能最佳。

5 结束语

本文基于得分归一化和系统融合的语音关键词检测方法，在连续语音识别解码过程中应用了关键

词相关的软 Beam 剪枝策略,如果关键词存在于语音中,即使得分较低也不会因为剪枝而丢失,因而有效降低了漏警率;并且在系统融合前,应用了得分归一化方法,使得各个子系统的候选检测结果的置信度得分在同一范围,具有可比性,最终合并结果的置信度得分更加准确;在实验中比较了常用得分归一化方法以及分数融合方法相比本文提出方法性能的优劣。本文方法充分利用了各个识别系统的互补性,相比传统系统融合方法考虑相对全面,步骤更加完善,虽然计算量相对增大,但是关键词检测性能得到了一定程度的提升。实验表明经过 STO 归一化处理后,关键词检测性能相比基线系统平均相对提升 30%,本文提出的系统融合方法相比最佳单一关键词检测系统关键词检测性能相对提升了 10%。

参考文献:

[1] 王炳锡, 屈丹, 彭煊. 实用语音识别基础[M]. 北京:国防工业出版社, 2005:287-291.
Wang Bingxi, Qu Dan, Peng Xuan. Practical fundamentals of speech recognition[M]. Beijing: National Defense Industry Press, 2005:287-291.

[2] Yu P, Seide F T B. A hybrid word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech[C]//Proceedings of the 5th Annual Conference of the International Speech Communication Association. Jeju Island, Korea: ISCA, 2004: 293-296.

[3] Chen B. Voice retrieval of Mandarin broadcast news speech[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2006, 20(1): 91-109.

[4] 孟猛, 王晓瑞, 梁家恩, 等. 一种基于互补声学模型的多系统融合语音关键词检测方法[J]. 自动化学报, 2009, 35(1): 39-45.
Meng Meng, Wang Xiaorui, Liang Jiaen, et al. A system combination based keyword-spotting method using complementary acoustic models[J]. Acta Automatic Sinica, 2009, 35(1): 39-45.

[5] Mangu L, Soltau H, Kuo H K, et al. Exploiting diversity for spoken term detection[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2013: 8282-8286.

[6] Mandal A, van Hout J, Tam Y C, et al. Strategies for high accuracy keyword detection in noisy channels[C]//Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon, France: ISCA, 2013:15-19.

[7] Zhang B, Schwartz R M, Tsakalidis S, et al. White listing and score normalization for keyword spotting of noisy speech [C]//Proceedings of the 13th Annual Conference of the International Speech Communication Association. Portland Oregon: ISCA, 2012:1832-1835.

[8] Allauzen C, Mohri M, Saraclar M. General indexation of weighted automata: Application to spoken utterance retrieval[C]//Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval. Stroudsburg, PA, USA: ACL, 2004: 33-40.

[9] Mohri M, Pereira F, Riley M. Weighted finite-state transducers in speech recognition [J]. Computer Speech and Language, 2002,16(1): 69-88.

[10] Can D, Saraclar M. Lattice indexing for spoken term detection[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(8): 2338-2347.

[11] Lee J H. Analyses of multiple evidence combination[C]//Proceedings of Workshop on Searching Spontaneous Conversational Speech. Philadelphia, PA, USA: ACM, 1997, 31(S1): 267-276.

[12] Meng S, Zhang W, Liu J. Combining Chinese spoken term detection systems via side-information conditioned linear logistic regression[C]//Proceedings of the 11th Annual Conference of the International Speech Communication Association. Chiba, Japan: ISCA, 2010: 685-688.

[13] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit[C]//Proceedings of the workshop on Automatic Speech Recognition and Understanding. Hawaii, USA: IEEE, 2011: 1-4.

[14] Ortmanns S, Ney H, Aubert X. A word graph algorithm for large vocabulary continuous speech recognition[J]. Computer Speech and Language, 1997, 11(1): 43-72.

作者简介:

李　鹏(1989-),男,硕士研究生,研究方向:语音信号处理、关键词检测,E-mail:15137172798@163.com。

屈　丹(1974-),女,副教授,研究方向:语音信号处理、模式识别,E-mail:qudanqudan@sina.com。