采用词图相交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：针对词图合并方法产生的词图冗余信息过多，规模较大，导致检索速度较慢的问题，本文提出了一种基于词图相交融合的语音关键词检测方法。首先，将不同语音识别系统产生的词图取交集，并对相同路径上的声学模型、语言模型得分进行得分融合；然后，对于融合后词图中存在的间断路径，直接利用性能最优的语音识别系统产生的词图进行补充，得到完整的融合词图；最后，在相交融合后的词图上进行关键词检测。实验表明，相交融合后的词图综合利用了各词图的得分信息，在基本不损失词图对正确内容覆盖率基础上，减少了冗余信息，有效降低了索引规模；并且在关键词检测性能 ATWV 指标下，基于词图相交融合的关键词检测方法相比词图合并方法相对提升 5.3%。

关键词：子空间高斯混合模型；深层神经网络；相交词图；关键词检测
中图分类号：TN911.7 **文献标识码**：A **文章编号**：1003-0530(2015)06-0702-08

Keyword Spotting Using the Lattice Intersection Fusion

LI Peng QU Dan
(Institute of Information Systems Engineering, PLA Information Engineering University, Zhengzhou 450001, China)

Abstract: In modern keyword spotting implementation, lattice combining is a useful method. Regrettably, it can result in massive redundant information which leads to a slow indexing. This paper propose a keyword spotting method based on lattice intersecting. Firstly, we take the intersection of two lattices, and combining the acoustic score and language score on the same path. Secondly, in order to solve the problem of gaps in the intersection lattice, we use the best single lattice to supplement it. Finally, the intersection lattice will be used for retrieval. With intersection lattice, it utilized the scores of each sub lattice. Without degrading the lattice coverage accuracy, it substantially eliminated redundant information and decreased indexing volume. The experimental results show that keyword spotting can improve 5.3% quality compared to lattice combining referred to keyword spotting ATWV indicator.

Key words: subspace Gaussian mixture model; deep neural network; intersection lattice; keyword spotting

1 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以互联网为载体，传播广泛，我们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 Spotting, KWS)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。

语音关键词检测系统首先应用大词表连续语音识别(Large Vocabulary Continuous Speech Recognition, LVCSR)系统将语音信号转换为文本形式，然后在文本上搜索用户提出的查询项。使用不同的语音识别系统进行识别，综合利用各个识别系统的知识能够提升检测系统的性能[1]。在此基础上，研究人员提出了基于融合的语音关键词检测技术，这项技术在检测系统的前端采用多套语音识别系统，并通过融合方法来提高整体系统的性能。

目前，主流的融合方法有检测结果层的融合和

收稿日期：2014-11-06；修回日期：2015-02-01
基金项目：国家自然科学基金(61175017)

识别结果层的融合。检测结果层的融合,是指对多套语音识别系统的识别结果分别建立索引,在各自得到检测结果后进行融合的方式[2-6]。但是这种方法的融合策略以及融合结果的置信度计算方法相对单一,融合性能提升有限。识别结果层的融合,是指直接融合多套语音识别系统的识别结果,然后在融合后的识别结果上建立索引,再进行关键词检测的技术。识别结果层的融合,可以显著地改进识别系统的性能和稳健性[7-8],进而提高关键词检测系统的性能。因此,识别结果层的融合成为目前语音关键词检测技术研究的热点之一。文献[9]将不同连续语音识别系统的首选结果(1-Best)组合成类似混淆网络(Confusion Network, CN)[10]的格式,并应用基于编辑距离的动态时间归整(Dynamic Time Warping, DTW)算法实现检索,一定程度上提高了关键词的检测性能,但是 1-Best 结果融合的关键词检测系统召回率较低,检测结果时间信息不够准确;为此 Dovey 等人[11]将连续语音识别的中间结果词图(Lattice)直接进行合并,它是增加一个新的初始节点,并将两个词图的起始节点连接到该初始节点,从而将两者的候选路径统一到一个拓扑结构中。实验证明,在合并后的词图上进行关键词检测,检测系统的性能有了明显的提升。但是,直接合并后的词图冗余信息过多,规模较大,因此检索速度较慢。

为了解决词图合并方法存在的问题,本文提出了一种基于词图相交融合的语音关键词检测方法。由于不同连续语音识别系统产生的词图的路径(候选结果)大部分相同,词图相交融合方法将不同语音识别系统产生的词图取交集,即保留词图中的相同路径,并对相同路径上的声学模型、语言模型得分进行得分融合。对于融合后词图中存在的间断路径,直接利用性能最优的语音识别系统产生的词图进行补充,最终得到完整的融合词图,然后在相交融合后的词图上进行关键词检测。本文词图相交融合的结果,综合利用了各词图的得分信息,并实现了减少冗余信息,降低索引规模的目的。实验部分对词图相交融合前后的性能进行了比较,验证了词图相交融合方法的有效性;并通过对比不同词图融合方法的关键词检测性能,验证了词图相交融合方法对检测性能提升的优越性。

本文的组织如下:第 2 节对词图合并方法进行了介绍;第 3 节介绍了基于加权有限状态机的词图相交算法;第 4 节介绍实验设置及实验结果;最后为结论部分。

2　词图合并方法

由于关键词检测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,因此常使用词图等多候选识别结果建立索引。在语音关键词检测中,词图的引入更是极大地提高了系统的召回率和检索性能。词图合并方法的提出[11],主要是为了对多个词图的信息进行综合利用,使其互相补充,从而提高连续语音识别系统的识别率,进而提升关键词检测的性能。本节简要介绍词图及词图合并方法。

2.1　词图(Lattice)

词图是 LVCSR 系统的中间识别结果,它不仅包含更多候选词识别结果,还记录了各候选识别结果的上下文路径、声学模型与语言模型得分以及时间点等信息[12]。词图是由解码器搜索解码网络,记录下搜索路径转化而成的,它的特性与连续语音识别系统的解码算法直接相关。Povey 等人[13]提出了一种基于加权有限状态机(Weighted Finite State Transducer, WFST)的词图生成算法,并应用在开源项目 Kaldi[14]中。这种算法产生的是一种隐马尔科夫(Hidden Markov Model, HMM)状态级词图,词图中没有重复的词序列路径信息,并且词图中的得分、时间信息相对准确。在基于 WFST 的词图生成算法中,应用有限状态机(Finite state Transducer, FST)表示词图。它的权重包含语言模型得分(graph cost)和声学模型得分(acoustic cost),状态转移中输入为 HMM 状态转移序列,输出为解码器的解码单元(词或音素,本文后续实验中为词)。如图 1 为"高州市"的词级词图结构图。

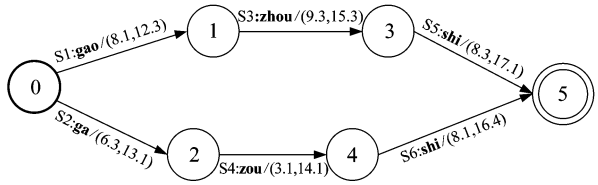

图 1　"高州市"词级词图结构图

Fig. 1　An example of Lattice structure

词图中 0 为初始状态,用加粗圆圈表示;5 为结束状态,用双圆圈表示。(x:y(w_1,w_2))表示状态间的转换关系,其中 x 表示输入,y 表示输出,(w_1,w_2)为权重信息。例如,从状态 0 到状态 1 的转换中,$s1$ 为输入 HMM 状态转移序列,"gao"为输出符号,权重信息(8.1,12.3)分别为语言模型得分与声学模型得分。

2.2 词图合并

以两个词图合并为例,它是增加一个新的初始节点,并将两个词图的起始节点连接到该初始节点,从而把两者的候选路径统一到一个拓扑结构中,词图合并方法如下图 2 所示:

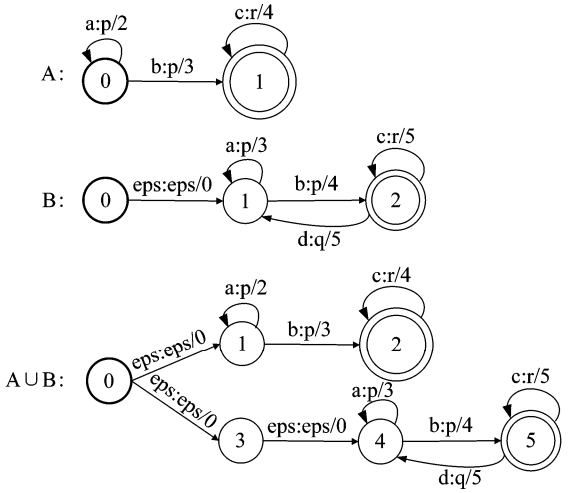

图 2 词图合并方法

Fig. 2 Lattice combination method

图 2 中 A 表示词图 L_1,B 表示词图 L_2,$A\cup B$ 为词图 L_1 与 L_2 合并的结果。为了简化本文将状态间的转移关系表示为(x:y/w),其中 x 表示输入,y 表示输出,w 为权重信息,eps 表示空符号。例如 A 中状态 0 到状态 1 的转换中,输入为符号 b,输出为符号 p,权重为 3。合并过程为首先建立初始节点 0,并将词图 L_1 与词图 L_2 的原起始节点连接到该初始节点,状态转移关系都设为(eps:eps/0),其余状态转移关系保持不变;然后依次改变状态标号,最后得到的合并词图中上部分为原词图 L_1,下部分为原词图 L_2,只是状态标号发生改变,状态转移关系保持不变。从合并词图可以直观看出,它对正确内容的覆盖率提升,但是合并后的词图冗余信息过多,规模较大。

3 基于 WFST 的词图相交算法

为了实现减少冗余信息,降低索引规模目的,本文提出了基于 WFST 的词图相交算法。在基于 WFST 的词图生成算法中,应用 FST 表示词图,其中状态转移的输出为解码器的解码单元(词或音素,本文后续实验中为词)。不同连续语音识别系统产生的词图中有很多路径相同,即状态转移的输出符号相同。词图相交算法的核心是将不同词图中存在相同输出符号的状态转移进行相交融合,融合为一个状态转移,并对权重信息进行分数融合。为了实现词图的相交融合处理,我们借鉴了 WFST 中的合成算法[15]的思想。本节主要介绍 WFST 中的合成算法及词图相交融合的实现步骤。

3.1 WFST 中的合成算法

WFST 中的合成算法是根据一个串联的 WFST 模型生成单一的 WFST,得到新 WFST 的输入输出关系与原串联 WFST 相同。以两个 WFST 合成为例,假设有 $T_1=(\Sigma_1,\Delta_1,Q_1,I_1,F_1,E_1,\lambda_1,\rho_1)$ 和 $T_2=(\Sigma_2,\Delta_2,Q_2,I_2,F_2,E_2,\lambda_2,\rho_2)$ 满足 $\Delta_1\subseteq\Sigma_2$,即第一个转换器的输出符号序列可作第二个转换器的输入符号序列。其中,Σ 表示有限输入字母表,Δ 表示有限输出字母表,Q 表示有限状态集合,$I\subseteq Q$ 表示初始状态集合,$F\subseteq Q$ 表示终止状态集合,$E\subseteq Q\times(\Sigma\cup\{\varepsilon\})\times(\Delta\cup\{\varepsilon\})\times K\times Q$ 为有限弧集合,表示状态转移,λ 和 ρ 分别为初始和终止权重函数。给定输入输出符号序列对 $(x,y)\in\Sigma_1^*\times\Delta_2^*$ 时,其中 Σ^* 为 Σ 的星闭包,又称克林闭包,表示由 Σ 中的符号组成的序列的集合以及空序列 ε,$\Sigma_1^*\times\Delta_2^*$ 描述了状态转移关系;则 T_1、T_2 串联的系统域可表示为如下形式:

$$[T_1\circ T_2](x,y)=\oplus_{z\in\Delta_1^*}[T_1](x,z)\otimes[T_2](z,y) \tag{1}$$

其中 $[T_1\circ T_2](x,y)$ 表示合成后的 WFST,输入为 x,输出为 y;$[T_1](x,z)$ 与 $[T_2](z,y)$ 分别表示 WFST 对应的输入输出关系;$\otimes$ 与 $\oplus$ 表示半环中的二元运算符,用于计算 WFST 的输出得分。式(1)表示首先将 x 输入 T_1,得到输出 z 与相应得分后,再将 z 作为 T_2 的输入,并得到输出 y 与相应得分。如果 WFST 所在半环对 $\otimes$ 运算满足交换率,则可以构建一

个新的 WFST,记为 T,并且 $\forall(x,y)\in\sum_1^*\times\Delta_2^*$,$[T](x,y)=[T_1\circ T_2](x,y)$,即为合成算法。合成算法是对 WFST 中的状态转移进行操作,设 T_1 中有状态转移(p_1,a,b,w_1,q_1),T_2 中有状态转移(p_2,b,c,w_2,q_2),其中(p_1,a,b,w_1,q_1)表示 T_1 中状态 p_1 到状态 q_1 的转移中输入符号为 a,输出符号为 b,权重信息为 w_1,T_2 中状态转移表示方式相同。T_1 中的状态转移输出符号为 b,T_2 中的状态转移输入符号也为 b,则合成结果 T 中可添加状态转移:

$$((p_1,p_2),a,c,w_1\otimes w_2,(q_1,q_2)) \qquad (2)$$

其中,(p_1,p_2)与(q_1,q_2)表示 T 中的状态。

3.2　词图相交算法

以两个词图相交融合为例,不同连续语音识别系统分别生成两个词图 L_1 与 L_2,假设词图 L_1 和 L_2 中分别存在状态转移(x_1,a,c,w_1,y_1)以及(x_2,b,c,w_2,y_2)。其中,两个词图中状态转移输出符号都为 c,w_1,w_2 表示权重信息,分别包含语言模型得分和声学模型得分。则基于 WFST 的词图相交算法具体步骤如下:

步骤 1　将词图 L_2 中所有的状态转移输出符号映射到输入符号,即将状态转移(x_2,b,c,w_2,y_2)变为(x_2,c,c,w_2,y_2),词图 L_2 变为 L_2';

步骤 2　将词图 L_1 与词图 L_2'中的状态转移进行合成算法处理,即状态转移(x_1,a,c,w_1,y_1)与(x_2,c,c,w_2,y_2)的合成结果 L 中可添加状态转移$((x_1,x_2),a,c,w_1\otimes w_2,(y_1,y_2))$,实现了存在相同输出符号的状态转移的相交融合,并且相交词图 L 保留词图 L_1 中的时间点信息;

步骤 3　若词图 L_1 与 L_2'中存在没有相同输出符号的状态转移,则合成结果为空,即相交后的词图 L 中存在间断的状态转移。将合成结果词图 L 与词图 L_1 进行比较,缺少的状态转移直接应用 L_1 补充,得到完整的相交词图。

本文在词图相交算法中,将性能最优的语音识别系统产生的词图作为 L_1。算法中相交词图 L 保留了词图 L_1 中状态转移的时间信息,即相交词图 L 与词图 L_1 中存在相同输出的状态转移的时间点信息相同。因此,本文将相交词图 L 与词图 L_1 进行比较,缺少的路径直接应用 L_1 补充,可以得到完整的相交词图,解决了相交词图存在间断路径的问题。具有相同输出符号的状态转移进行相交融合,需要对声学模型、语言模型得分进行得分融合,本文采用加权相加的方法,将可以相交融合的状态转移中的声学模型、语言模型得分赋予一定权值,权值和为 1,再将相加结果赋予相交后的词图。词图相交方法如图 3 所示。

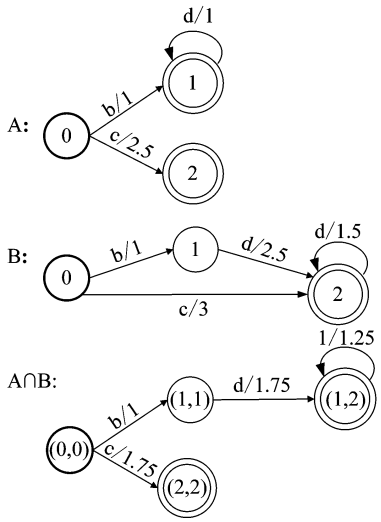

图 3　基于 WFST 的词图相交方法

Fig. 3　Lattice intersection method based on WFST

图 3 中 A 表示词图 L_1,B 表示词图 L_2,$A\cap B$ 为词图 L_1 与 L_2 相交的结果,为了直观的表示词图相交算法状态转移中输出符号的变化,图中状态间的转移关系表示为(y/w),其中 y 表示输出,w 为权重信息。从图 3 中可以看出,L_1 中存在一条输出符号序列为“bd”的有效路径,状态转移为$(0\rightarrow1\rightarrow1)$;$L_2$ 中也存在一条输出符号序列“bd”的有效路径,状态转移为$(0\rightarrow1\rightarrow2)$;则经过基于 WFST 词图相交算法处理后,$A\cap B$ 中得到一条有效路径,输出符号序列为“bd”,状态转移为$(0,0)\rightarrow(1,1)\rightarrow(1,2)$;权重为 L_1 和 L_2 路径上所有权重进行⊗运算的结果。图中词图 L_1、L_2 中的权重信息分别赋予权值为 0.5,再将加权相加后的结果赋予相交后的词图。可以直观看出相交处理后的词图冗余信息减少,规模较小,并包含了所有具有相同输出符号的状态转移关系。

4　实验结果及分析

4.1　实验语料库及实验设置

本文实验采用微软汉语语料库,其中训练集由 100 个男性录音组成,每人大约 200 句,共 19688 句话,454315 个音节,约 33 小时的语音数据。测试集

包含25个男性录音,每人20句,共500句话,每句话大约5s时长。关键词查询项从测试集的标注中,依据测试集的词频选取。共选取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。

识别结果层的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。我们利用开源Kaldi工具箱搭建了两套声学模型:基于子空间高斯混合模型(Subspace Gaussian Mixture Model, SGMM)[16]和基于深层神经网络(deep neural network, DNN)[17]声学模型。传统的GMM模型,应用EM(Expectation maximization)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。然而它没有考虑如何更好的区分各类别,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),并应用了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。

实验设置中,SGMM连续语音识别系统应用13维MFCC(Mel frequency cepstrum coefficient)特征及其一、二阶差分,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)[18]方法对每一个说话人的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1935个不同的绑定状态(tied states),应用Kaldi工具箱训练基于子状态的扩展SGMM声学模型。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1935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(Linear Discriminant Analysis, LDA)[19],最大似然线性变换(Maximum Likelihood Linear Transform, MLLT)[20],和特征域最大似然线性变换(Feature-space Maximum Likelihood Linear Regression, FMLLR)[21]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。本文均采用无调音节进行解码识别,并生成词级词图。最后利用Kaldi工具包进行基于WFST的关键词检测[22]。

4.2 词图相交中得分融合权重对识别性能的影响

词图相交融合中,需要对相同路径位置上的声学模型、语言模型得分进行得分融合。权重的大小会直接影响基于相交词图的二次解码的连续语音识别错误率(WER),间接影响基于相交词图的关键词检测性能。实验中通过观察不同权重α对基于相交词图的二次解码的连续语音的词错误率的影响,选取最佳得分规整权重。其中L_1表示DNN系统生成的词图,它的得分融合权重为α;L_2表示SGMM系统生成的词图,它的得分融合权重为$1-\alpha$。

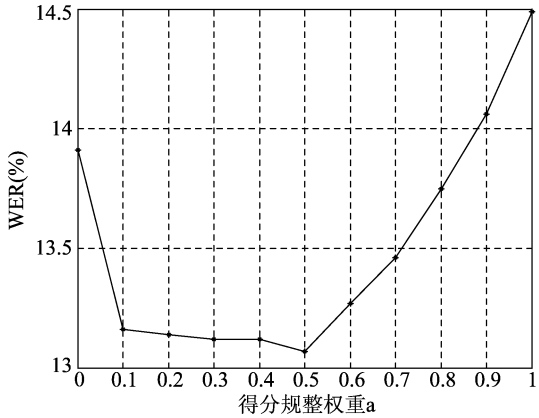

图4 得分融合权重α对相交词图识别性能的影响

Fig. 4 WER according to different score combination weight

从图4中可以看出,当得分权重从$\alpha=0$开始增大时,词错误率降低,这是由于L_1中的得分信息对于相交词图二次解码的贡献开始增大,使得最终相交词图的得分更准确。当得分权重$\alpha=0.5$条件下,基于相交词图的二次解码的连续语音的词错误率最小为13.07%。当得分权重α进一步增大时,词错误率增大,识别性能降低,这说明在对原始词图

的声学模型和语言模型得分进行分数融合时,需要选择合适的权重,不可偏废一方,这样才能使得基于相交词图的二次解码连续语音识别词错误率最低。

4.3　词图融合前后的性能比较

为了比较本文词图相交融合方法的效果,我们建立了三套基线系统和一套关于本方法的新系统:

1)DNN:直接利用 DNN 连续语音识别系统生成词图 L_1,并在词图上进行关键词检测实验;

2)SGMM:直接利用 SGMM 连续语音识别系统生成词图 L_2,并在词图上进行关键词检测实验;

3)词图合并方法:将 DNN 与 SGMM 连续语音识别系统生成的词图 L_1 与 L_2 合并,并在合并词图上进行关键词检测实验;

4)词图相交融合方法:即本文提出的新方法,将 DNN 与 SGMM 连续语音识别系统生成的词图 L_1 与 L_2 相交融合,并在相交融合后的词图上进行关键词检测实验。

本文用图错误率(graph error rate,GER)和词图密度(Lattice density)来评价词图融合前后的性能。图错误率(GER)定义为:

$$\mathrm{GER}(L)=\min_{w_1^N\in L}\frac{Lev(w_1^N,\tilde{w}_1^{N_r})}{N_r} \tag{3}$$

式中 $Lev(w_1^N,\tilde{w}_1^{N_r})$ 表示词图中词序列 w_1^N 与参考标注序列 $\tilde{w}_1^{N_r}$ 的 Levenshtein 距离,N_r 为参考标注中的词个数。图错误率表示了词图中和参考标注序列最相似的识别候选的错误率。

词图密度是指词图 L 中弧数目 $|E(L)|$ 与对应参考标注序列中词个数的比值。公式如下:

$$\mathrm{density}(L)=\frac{|E(L)|}{N_r} \tag{4}$$

词图密度表示平均一个参考标注需要多少个弧(即候选标注)来表示,从词图密度中可以直观看出词图的规模大小。图错误率越低,词图密度越小,则词图性能越佳。

表 1 给出了词图融合前后的图错误率、词图密度以及基于词图的二次解码的连续语音识别词错误率(WER)。其中 L_1 表示 DNN 系统生成的词图,L_2 表示 SGMM 系统生成的词图。

表 1　词图融合前后的性能

Tab. 1　The performance of the different Lattice

不同词图	GER(%)	Lattice Density	WER(%)
L_1	10.93	6.4	13.91
L_2	11.84	9.8	14.49
L_1 与 L_2 合并	10.06	18.3	12.19
L_1 与 L_2 相交	10.52	7.1	13.07

从表 1 中可以看出 SGMM 与 DNN 系统生成的词图进行合并或相交融合后,GER 相比融合前都有所下降,合并词图的 GER 下降到 10.06%,相交词图的 GER 下降到 10.52%。由于合并词图将 L_1 与 L_2 合并到一个拓扑结构中,对连续语音识别正确内容的覆盖率增加,所以合并词图相比相交词图的 GER 略小。但是合并后词图规模急剧增大,其词图密度 18.3 明显大于相交词图的 7.1。实验结果显示相交词图相比合并词图,在基本没有损失对正确内容覆盖率的基础上,得到了更小规模的词图。

4.4　词图相交中关键词检测门限的选取

关键词检索门限,即为关键词检测系统设置一个门限 θ,当置信度得分大于门限值 θ,才被认为关键词检测结果正确。全局门限是最简单的阈值策略,即为所有的查询项设置相同的阈值 θ,接受满足置信度得分大于阈值 θ 的候选者。词图相交融合过程中对相同路径上的声学模型、语言模型得分进行了得分融合,会直接影响关键词检测结果的置信度得分。为了得到最佳阈值 θ,本文研究了不同阈值下,基于相交词图的语音关键词检测方法的实际查询词权重代价(Actual Term-Weighted Value,ATWV)[23] 得分。最佳阈值 θ 下 ATWV 得分最高。

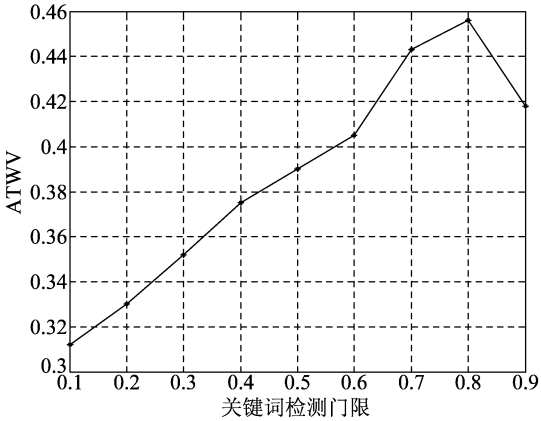

图 5　关键词检测门限对 ATWV 的影响

Fig. 5　ATWV results according to different spoken term detection threshold

从图5中可以看出,随着关键词检索门限 θ 的增大,关键词检测性能 ATWV 得分增大。这是由于 ATWV 得分与漏警率、虚警率直接相关,只有在漏警率与虚警率都相对较低的情况下 ATWV 得分才会较高。提高关键词检测门限,在漏警率基本不变的情况下明显降低了系统的虚警率,所以 ATWV 提升。当阈值 $\theta=0.8$ 时,ATWV 得分最高为 0.456。当阈值 θ 进一步增大时,漏警率增大,ATWV 性能降低,所以最佳的阈值 $\theta=0.8$。

4.5 词图融合前后对关键词检测性能 ATWV 的影响

为了检验本文提出的词图相交方法的关键词检索性能,分别以4.3节建立的四套系统生成的词图作为关键词检索系统的输入进行检索,得到不同系统的关键词检测性能 ATWV 得分。

表2 词图融合前后对关键词检测性能 ATWV 的影响

Tab.2 ATWV relative improvement as a function of different Lattice

不同词图	ATWV	相比 L_1 相对提升	相比合并相对提升
L_1	0.412	-	-
L_2	0.407	-	-
L_1 与 L_2 合并	0.433	6.4%	-
L_1 与 L_2 相交	0.456	12.1%	5.3%

从表2中可以看出相交词图关键词检索的 ATWV 得分最高为 0.456,相比最佳单系统 L_1 相对提升 12.1%,相比 L_1 与 L_2 合并方法相对提升 5.3%。基于相交词图的连续语音识别率错误率 13.07% 高于词图合并方法的 12.19%,但是其关键词检测性能却优于词图合并方法。原因在于语音关键词检索中,对关键词检出结果的可靠程度给出一个置信度分数,用其评价检出关键词的可信度。目前,大多数语音关键词检索系统应用关键词在词图中得到的后验概率作为其置信度得分来进行关键词确认。用于置信度计算的特征主要包括声学模型得分、语言模型得分、词候选驻留时间等识别结果本身的信息。相交融合后的词图有效融合 L_1 与 L_2 的得分信息,置信度分数更加准确,关键词检测性能 ATWV 得分更高。实验表明相交词图规模较小,索引规模降低,并且在基本不损失词图对正确内容覆盖率的基础上,获得了更好的关键词检测性能。

5 结论

本文针对合并词图冗余信息过多,规模较大,导致检索速度较慢的问题,提出了一种基于词图相交融合的语音关键词检测方法。该方法将不同语音识别系统产生的词图取交集,并对相同路径上的声学模型、语言模型得分进行得分融合。对于融合后词图中存在的间断路径,直接利用性能最优的语音识别系统产生的词图进行补充,最终得到完整的融合词图。实验表明相交融合后的词图在基本不损失对正确内容覆盖率基础上,有效减小了索引规模,从而加快了关键词检索速度,并且在关键词检测性能 ATWV 指标下,基于词图相交融合的关键词检测方法性能优于词图合并方法。

参考文献

[1] 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(ICASSP). Florence, Italy: IEEE, 2013: 8282-8286.

[2] Abad A, Rodríguez-Fuentes L J, Penagarikano M, et al. On the calibration and fusion of heterogeneous spoken term detection systems[C]//Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH). Lyon, France: ISCA, 2013: 20-24.

[3] Mamou J, Cui J, Cui X, et al. System combination and score normalization for spoken term detection[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing(ICASSP). Florence, Italy: IEEE, 2013: 8272-8276.

[4] Akbacak M, Burget L, Wang W, et al. Rich system combination for keyword spotting in noisy and acoustically heterogeneous audio streams[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing(ICASSP). Florence, Italy: IEEE, 2013: 8267-8271.

[5] Van Hout J, Mitra V, Lei Y, et al. Recent improvements in SRI's Keyword Detection System for Noisy Audio[C]//Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH). Singapore: ISCA, 2014:1727-1731.

[6] Chiu J, Wang Y, Trmal J, et al. Combination of FST and CN search in spoken term detection[C]//Proceed-

ings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH). Singapore: ISCA, 2014:2784-2788.

[7] Fiscus J G. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)[C]//Proceedings of the Workshop on Automatic Speech Recognition and Understanding (ASRU). Santa Barbara, CA, USA: IEEE, 1997: 347-354.

[8] Hoffmeister B, Klein T, Schlüter R, et al. Frame based system combination and a comparison with weighted ROVER and CNC[C]//Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH). Pittsburgh, PA, USA: ISCA, 2006:1523-1525.

[9] Natori S, Furuya Y, Nishizaki H, et al. Spoken Term Detection Using Phoneme Transition Network from Multiple Speech Recognizers' Outputs[J]. Information and Media Technologies, 2013, 8(2): 457-466.

[10] Minescu B, Damnati G, Béchet F, et al. Conditional use of word lattices, confusion networks and 1-best string hypotheses in a sequential interpretation strategy[C]//Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH). Antwerp, Belgium,2007: 1617-1620.

[11] Xu H, Povey D, Mangu L, et al. Minimum Bayes Risk decoding and system combination based on a recursion for edit distance[J]. ISCA Computer Speech and Language, 2011, 25(4): 802-828.

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

[13] Povey D, Hannemann M, Boulianne G, et al. Generating exact lattices in the WFST framework[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto, Japan: IEEE, 2012: 4213-4216.

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

[15] Mohri M, Pereira F C N, and Riley M. Speech Recognition with Weighted Finite-State Transducers[M]. Handbook of Speech Processing, Verlag Berlin Heidelberg, Springer, 2008:559-582.

[16] Povey D, Burget L, Agarwal M, Akyazi P, et al. Subspace gaussian mixture models-a structured model for speech recognition[J]. ISCA Computer Speech and Language, 2011, 25(2): 404-439.

[17] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[18] Tibrewala S, Hermansky H. Multiband and adaptation approaches to robust speech recognition[C]//Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH). Rhodes, Greece: ISCA,1997:2619-2622.

[19] Kumar N, Andreou A G. Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition[D]. Johns Hopkins University, 1997.

[20] Gales M J F. Semi-tied covariance matrices for hidden Markov models[J]. IEEE Transactions on Speech and Audio Processing, 1999, 7(3): 272-281.

[21] Ghoshal A, Povey D, Agarwal M, et al. A novel estimation of feature-space MLLR for full-covariance models[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing(ICASSP). Dallas, Texas, USA: IEEE, 2010: 4310-4313.

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

[23] Fiscus J G, Ajot J, Garofolo J S, et al. Results of the 2006 spoken term detection evaluation[C]//Proceedings of Workshop on Searching Spontaneous Conversational Speech(SSCS). Amsterdam, Netherlands: ACM, 2007, 7: 51-57.

作者简介

李　鹏　男,1989 年生,陕西韩城人。信息工程大学信息系统工程学院研究生,2012 年毕业于信息工程大学信息系统工程学院并获学士学位。主要研究方向为语音关键词识别。

E-mail:15137172798@163.com

屈　丹　女,1974 年生,毕业于信息工程大学并获博士学位,信息工程大学信息系统工程学院副教授,主要研究方向为语音信号处理、模式识别。

E-mail:qudanqudan@sina.com