doi:10.3969/j.issn.1673-4785.2010.05.009

结合关键词混淆网络的关键词检出系统

张 磊,陈 晶,项学智,贾梅梅
(哈尔滨工程大学 信息与通信工程学院,黑龙江 哈尔滨 150001)

摘 要:为了高效地从大词汇量连续语音识别(LVCSR)的多候选中得到关键词结果,保证最小词错误率,提出了将混淆网络的思想应用到关键词检出系统中.在传统混淆网络生成方法基础上,提出一种改进的更加适合于关键词检出的关键词混淆网络作为关键词检出的中间结构,该方法只对所有关键词竞争候选生成带有得分标记的关键词混淆网络,突出候选之间竞争关系,并根据得分标记确定关键词.与传统的N-best作为中间结构的关键词检出系统比较,基于混淆网络的关键词检出系统的召回率为87.11%,提高了21.65%.实验表明,在提高召回率的同时,所提方法具有关键词直接定位的特点,因此具有较低的时间开销.

关键词:关键词检出;混淆网络;语音识别
中图分类号:TP391;TN912 **文献标识码**:A **文章编号**:1673-4785(2010)05-0432-04

Research of keyword spotting based on a keyword spotting confusion network

ZHANG Lei, CHEN Jing, XIANG Xue-zhi, JIA Mei-mei
(College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: In order to achieve a higher keyword recall rate from large vocabulary continuous speech recognition (LVCSR) and minimize the word error rate, a confusion network was used in a keyword spotting system. Moreover, an improved method of generating a keyword confusion network which was more suitable for keyword spotting was proposed based on the traditional algorithm. This method only focused on keyword competitions, and was capable of transforming all the keyword competitions into a confusion network with a marked score, and highlighted competitions to all the candidates. Compared with the traditional keyword spotting system which uses N-best as the medium structure, the proposed method increased the recall rate of confusion network to 87.11%; compared with the keyword spotting system based on N-best, there is a 21.65% improvement in the recall rate. Experiments show the proposed method could locate keywords directly, besides increasing the recall rate, so the system costs less time.

Keywords: keyword spotting; confusion network; speech recognition

目前,大词汇量连续语音识别技术已经取得了巨大的进展.然而,这并不能满足使用的需要.如何高效地管理、分类、查找这些大容量的音频文件成为语音识别研究领域的又一挑战[1-3].关键词检出(keyword spotting)的目的是从连续无限制的语音流中识别出给定的若干关键词,按其检出方式可分为基于连续语音识别和基于补白(fill)模型2种方式[4].目前,应用较多的是基于连续语音识别的关键词检出系统,主要是指对连续语音经过声学解码处理后检出关键词.N-best和Lattice是连续语音识别2种最常用的结果组织形式.N-best是将语音识别结果按后验概率由大到小排列,并选取出前N个识别结果.目前,传统的关键词检出系统多是基于N-best结构设计.Lattice可以提供足够的候选以保证检出正确,但需要高效的解码算法,难以实现.2000年Mangu[5]提出用混淆网络(confusion network)优化Lattice,随后,Xun等人[6]提出了一种快速的混淆网络生成算法,并取得良好效果.2007年Zhang等人[6]提出将混淆网络优化Lattice的方法应用在关键词检出中,但这种新的尝试是基于对整个Lattice进行混淆网络的转换,而后进行关键词检出,并没有将混淆网络与关键词结合,消耗时间巨大.本

收稿日期:2009-12-03.
基金项目:国家自然科学基金资助项目(60702053);黑龙江省青年骨干教师支持计划资助项目(1155G17).
通信作者:张 磊. E-mail:zhanglei@hrbeu.edu.cn.

文将关键词检出融合到混淆网络生成算法中,在简化混淆网络生成算法的同时,突出关键词候选之间的竞争关系,实验表明,该方法提高了关键词检出效率,且该算法时间复杂度低,易于实现.

1　系统框架

基于关键词混淆网络的关键词检出系统框架如图 1 所示.

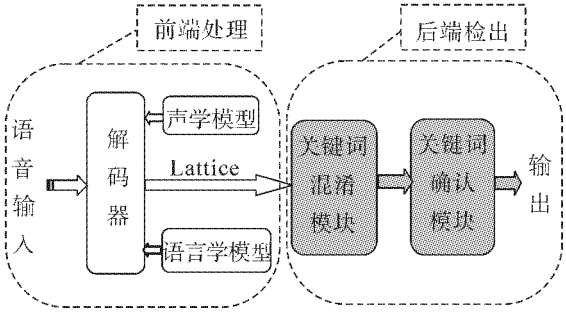


图 1　关键词检出系统框图

Fig. 1　Diagram of keyword spotting system

前端处理部分采用 HTK 工具箱进行模型的训练和识别[8]. 系统声学模型为上下文相关的三因素 tri-phone 模型,拓扑结构为带跳转的自左向右结构,每个模型取 5 个状态,按照字典拼接成音节模型进行识别;语言学模型为基于音节的 bi-gram 语言模型,并应用 Katz 算法[9]进行平滑.

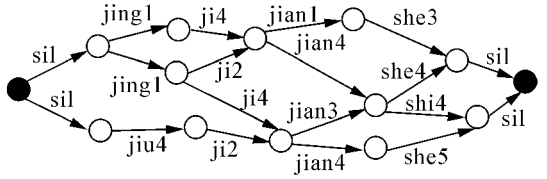
后端检出是本文研究的重点. 首先,把解码后的 Lattice 结构作为关键词混淆网络生成的输入,通过匹配关键词,在 Lattice 中生成带有得分标记的关键词混淆网络. 其次,要在生成的关键词混淆网络中对关键词进行确认. 对于 Lattice 中的每个弧都有标记的声学得分和语言学得分,生成混淆网络后,相应地转化为混淆网络中弧的得分,代表了识别结果与待识别语音的匹配程度. 最后,根据弧的得分可判断出最有可能的候选.

2　关键词混淆网络

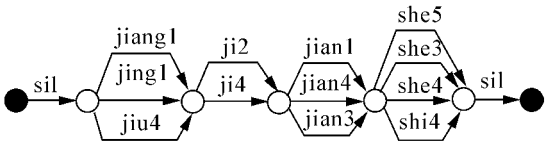
2.1　混淆网络的概念

混淆网络是 Lattice 中弧和节点通过动态对齐后生成的结构. 在这种结构中,所有竞争同一个发音位置的词形成一个集合,然后把这些集合按照时间顺序依次连接起来,在每个集合中挑选最可能的词形成最佳词串. 图 2 给出了 Lattice 和混淆网络对比的例子. 以“经济建设”为例,其 Lattice 结构如图 2(a)所示;其混淆网络结构如图 2(b)所示. 其中,节点的横向排列严格按照时间先后顺序.

在图 2 中,混淆网络结构很好地解决了 Lattice 中的识别结果在时间上相互交叠的现象. 在混淆网络中,同一语音单元的不同识别结果及其对应的得分体现地非常明显. 句子的识别结果只要在每个音节混淆网络中通过一定的约束条件找出最佳识别结果,连接即可. 它突出了识别结果之间的竞争关系,这也正为关键词检出提供了有利条件. 在对 Lattice 的解码过程中,人们常常采用基于句子的最大后验概率(MAP)准则的解码方法. 这种方法更加关注最小化句子错误率,不能保证最小化词错误率. 但在混淆网络的解码过程中,通过在每个候选集合中选择后验概率最大的词,可以对它进行高效的最小化词错误率解码操作,这样就保证了关键词部分的正确率.



(a)Lattice 结构



(b) 混淆网络结构

图 2　Lattice 和混淆网络对比示意图

Fig. 2　Diagram of Lattice and confusion network

2.2　关键词混淆网络的生成

关键词混淆网络是根据 Lattice 生成,为了区别表示 Lattice 和混淆网络中的节点和弧,Lattice 中的节点和弧用小写字母表示,其中节点为$\{n_0,n_1,\cdots\}$,每个节点 n_i 含有一个时间标记 $t(n_i)$,弧 $e_{u\to v}$ 表示节点 u 连接到节点 v 的一个弧. 混淆网络中的节点用大写字母表示,如$\{N_0,N_1,\cdots\}$,其中 N_i 如图 2(b)所示的白色节点,对应 Lattice 中的节点集合,$E_{N_i\to N_j}$ 表示混淆网络中节点 N_i 连接到节点 N_j 的一组弧.

对于混淆网络和 Lattice 节点之间的关系,有以下几点约束:

1)对于 Lattice 中节点 n_i 和 n_j,如果在混淆网络中属于一个集合 N_k,则时序关系决定序号关系,即 $t(n_i)<t(n_j)$ 可得 $i<j$;$t(n_i)=t(n_j)$ 可得 $i=j$.

2) $\forall e_{u\to v}\in E$,如果对于 $u\in N_i$ 和 $v\in N_j$,$e_{u\to v}$ 相当于一组 $E_{N_m\to N_n}$,则可以将 $e_{u\to v}$ 的 2 个端点对应到混淆网络中相邻的 2 个节点上,分别属于 N_m 和 N_n,其中 $n=m+1$,$i\leqslant m\leqslant n\leqslant j$.

在上述约束条件的基础上,可以进一步构建关键词混淆网络. 和传统的混淆网络相比,关键词混淆网络直接在 Lattice 中定位关键词,以关键词的第 1 个音节为切入点,通过判断相邻节点的连接情况,计算得到包含关键词的混淆网络,并且标注关键词竞争候选生成的得分标记. 具体算法改进如下:

1)把关键词转化为对应的音节串:$K_1K_2\cdots K_M$(M 为关键词音节数,这里以 $M=2$,关键词为“经济”为例,则对应音节串 K_1 为 jing1,K_2 为 ji4).

2)遍历 Lattice 中的所有节点,找到与 K_1 相匹配的节点 n_k,对应的混淆网络节点设为 N_{k_1}.

3)关键词混淆网络节点生成.

a)若节点 n_{k-1} 与集合 N_{k_1} 中包含的所有节点都没有弧连接,则称该节点与集合 N_{k_1} 之间无弧连接,否则为有弧连接.

b)将 n_k 前面的节点并与 N_{k_1} 中无弧连接的节点合并到混淆网络节点 N_{k_1} 中. 直到找到与 N_{k_1} 之间有弧连接的节点时,停止搜索.

c)将 n_k 后面的节点并与 N_{k_1} 无弧连接的节点合并到混淆网络节点 N_{k_1} 中. 直到找到与 N_{k_1} 之间有弧连接的节点时,停止搜索.

d)找 n_k 后面的节点与 N_{k_1} 有弧连接的第1个节点,作为 N_{k_2},按照上述原则 c)形成 N_{k_2} 集合. 如多字词,按照相同原则形成多个混淆网络集合.

4)关键词混淆网络的弧的生成.

对 Lattice 中每个弧 $e_{k\to n_i}$,n_k 属于 N_{k_1},n_i 属于 N_t. 当 $t=k_2$ 时,$e_{k\to n_i}$ 属于 $E_{N_{k_1}\to N_{k_2}}$;否则,当为多字词时,按下述原则判断该弧是否属于集合 $E_{N_{n-1}\to N_n}$,$k_1+1\leqslant n\leqslant k_2$. 其中 n 为

$$n=\arg\max_{s+1\leqslant k\leqslant t}\{\mathrm{sim}(E_{N_{k-1}\to N_k},e)\},$$

$$\mathrm{sim}(E_{N_{k-1}\to N_k},e)=\frac{1}{|E_{N_{k-1}\to N_k}|}\times\sum_{l\in E_{N_{k-1}\to N_k}}\mathrm{sim}(w(l),w(e))\mathrm{overlap}(E_{N_{k-1}\to N_k},e).$$

式中:$w(l)$ 和 $w(e)$ 为弧 l 和 e 对应的词;sim(·,·)是指2个词的语音相似度,这里采用编辑距离;overlap($E_{N_{k-1}\to N_k}$,e)是指弧 $E_{N_{k-1}\to N_k}$ 和 e 的归一化时间交叠.

以关键词“经济”为例,混淆网络如图3所示. 在图3中,关键词混淆网络的每一条弧上都标记了其对应的结果和对应的得分. 其中,a 为声学概率似然得分,l 为语音学概率似然得分. 基于生成的关键词混淆网络,进行关键词的确认.

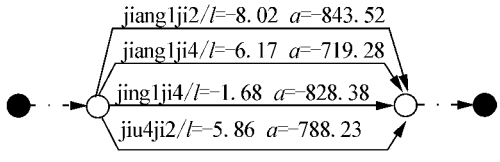


图3 关键词混淆网络示意图

Fig. 3 Diagram of keyword confusion network

2.3 关键词的确认

对于生成的关键词混淆网络,首先对每个候选的声学及语言学概率似然得分利用公式归一化.

$$y=\frac{x-\min(v)}{\max(v)-\min(v)}.$$

式中:v 为关键词混淆网络中声学或语言学得分值的集合;min(v)和 max(v)分别为集合中的最小和最大值;x 为待归一化得分;y 为归一化后的得分. 得分归一化后,再对声学及语言学概率得分赋以不同的权重,最后计算得分的和值,把它标记为关键词混淆网络中弧的得分. 以关键词“经济”为例,得分归一化并求和的结果如图4所示.

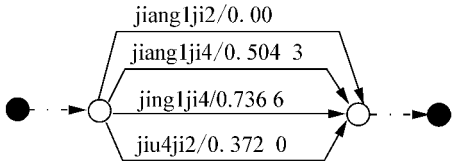


图4 关键词确认示意图

Fig. 4 Diagram of keyword verification

计算得分的和之后,在每一个混淆网络中选出得分最高的弧候选,判断是否为关键词. 如果是,输出关键词及时间标记;如不是,跳过并进入下一关键词混淆网络.

3 实验结果

3.1 实验环境和评价标准

实验利用 HTK 工具箱作为前期训练和识别. 训练语料为国家“863”语料库,测试语料选择500句话,在测试语料中随机选择中国、世界等20个二字词作为关键词,其中关键词,出现194次. 测试的性能评价标准包括召回率和误识率,召回率等于正确检出关键词数比关键词总数,误识率等于错误检出关键词数比检出关键词总数.

3.2 实验结果及分析

为了对比关键词混淆网络的关键词检出性能,实验基线系统将采用常用的 N-best 结果作为中间结构,并在其中查找关键词. N-best 中的 N 值取不同时,对检出结果也会有不同的影响. 如表1所示,给出了不同的 N 值对应的关键词召回率和误识率.

表1 N-best 实验结果比较

Table 1 The comparison of different N in N-best

N	召回率/%	误识率/%
1	56.19	2.68
10	61.34	4.03
20	65.46	3.79

如表1所示,当 N 取20时,关键词召回率及误识率明显优于 $N=1$ 和 $N=10$. 这是由于当候选增多时,原来未被检出的关键词有可能在增多的候选中被检出. 而3.79%的误识率也说明,当 N-best 中 N 取20时,可达到相对理想的检出性能.

对比实验在基线系统的基础上,将 Lattice 转化为关键词混淆网络后,在其中查找关键词. 其中,语言学得分权重设为0.7,声学得分0.3. 图5所示是基于关键词混淆网络系统的 ROC 曲线.

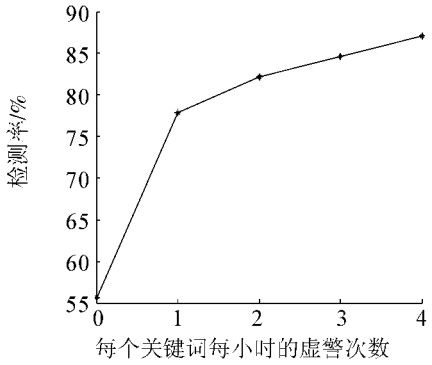


图 5　基于 CN 的关键词检出系统的 ROC

Fig. 5　ROC of keyword spotting based on CN

对于 2 种不同的方法,实验结果对比如表 2 所示. 相对于 20-best 的 65.46% 的结果,在关键词混淆网络中检出关键词的召回率提高了 21.65%,这是由于关键词混淆网络方法同时考虑了声学得分和语言学得分,突出关键词部分的竞争,并且最小化词错误率,所以在混淆网络结构中检出关键词比直接在 N-best 结果中检出关键词的能力明显提高;但是误识率却上升 5.35%,这是因为随着检出关键词数的增加,误识的关键词数也会上升.

表 2　实验结果比较

Table 2　The comparison of different methods　　%

实验方法	召回率	误识率
20-best	65.46	3.79
关键词混淆网络	87.11	9.14

4　结束语

本文将关键词检出技术融合到混淆网络生成中,从关键词的首音节开始只生成和关键词相关的部分混淆网络,因此和传统的混淆网络生成算法相比,具有速度快、定位准的优点. 同时利用归一化后验概率进行确认,和 N-best 结构的关键词检出系统相比,具有较高的检出率. 下一步工作将考虑词表进一步扩大到包含三字词和四字词,并在关键词混淆网络生成中,考虑加重语言学模型的概率信息. 在实验过程中,发现误识率增加是由于音调的误识造成,因此下一步工作将在关键词混淆网络中,将具有相同音节不同音调的各竞争候选整合,来弥补音调误识带来的影响.

参考文献:

[1]叶靓,王智斌,邵谦明. 基于相关反馈的语音检索引擎[J]. 计算机工程, 2007, 33(17): 228-230.
YE Liang, WANG Zhibin, SHAO Qianming. Speech retrieval engine based on relevance feedback[J]. Computer Engineering, 2007, 33(17): 228-230.

[2]王让定,袁旭海,徐霁. 一种新颖的混合语音检索算法[J]. 计算机应用研究, 2008, 25(5): 1349-1351.
WANG Rangding, YUAN Xuhai, XU Ji. Novel mixing speech retrieval algorithm[J]. Application Research of Computers, 2008, 25(5): 1349-1351.

[3]陈立伟,宋宪晨,章东华,等. 一种基于优化神经网络的语音识别[J]. 应用科技, 2008, 35(2): 17-20.
CHEN Liwei, SONG Xianchen, ZHANG Dongsheng, et al. Speech recognition using an optimized wavelet neural network[J]. Applied Science and Technology, 2008, 35(2): 17-20.

[4]郑铁然,韩纪庆. 汉语语音检索中基于音节的索引方法研究[C]//第八届全国人机语音通讯学术会议论文集. 北京, 中国, 2005: 419-424.
ZHENG Tieran, HAN Jiqing. Study on syllable based indexing methods in mandarin speech retrieval[C]//Proceedings of National Conference on Man-Machine Speech Communication. Beijing, China, 2005: 419-424.

[5]MANGU L, BRILL E, STOLCKE A. Finding consensus in speech recognition: word error minimization and other applications of confusion networks[J]. Computer Speech and Language, 2000, 14(4): 373-400.

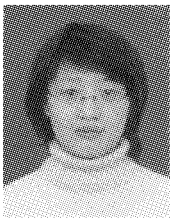
[6]XUE Jian, ZHAO Yunxin. Improved confusion network algorithm and shortest path search from word Lattice[C]// Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Philadelphia, USA, 2005: 853-856.

[7]ZHANG Pengyuan, SHAO Jian, ZHAO Qingwei, et al. Keyword spotting based on syllable confusion network[C]//The Third International Conference on Natural Computing. Haikou, China, 2007: 656-659.

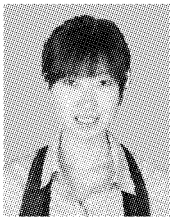
[8]YONG S, EVERMANN G, GALES M. The HTK book(for HTK 3.3)[EB/OL]. [2009-11-25]. Http://htk.eng.cam.ac.uk.

[9]GOODMAN J T. A bit of progress in language modeling[J]. Computer Speech and Language, 2001, 15(4): 403-434.

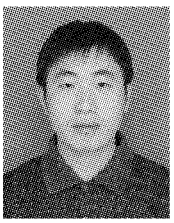
作者简介:



张　磊,女,1973 年生,副教授,主要研究方向为语音信号处理,承担或参与 4 项国家自然科学基金项目,发表学术论文 30 余篇.



陈　晶,女,1984 年生,硕士研究生,主要研究方向为语音信号处理.



项学智,男,1979 年生,讲师、博士,主要研究方向为信号处理,参与多项国家自然科学基金项目,发表学术论文 20 余篇.