文章编号:1004-9037(2014)02-0280-06

融合查询扩展和动态匹配的集外词检测

郑永军 张连海

(解放军信息工程大学信息系统工程学院,郑州,450001)

摘要:针对关键词检测中的集外词问题,提出了一种融合查询扩展和动态匹配的方法。查询扩展和动态匹配是在不同的层面补偿集外词发音的不确定性。考虑到两者潜在的互补性,采用两种融合方法:一种方法是结果融合,分别应用查询扩展和动态匹配并行的检测集外词,然后合并检测结果;另一种是置信度融合,融合最小编辑距离和发音得分构成混合置信度进行集外词的检出与确认。实验结果表明,第二种融合方法的效果更好,品质因数相比基线系统有了显著提升。

关键词:关键词检测;查询扩展;动态匹配;集外词

中图分类号:TP391 **文献标志码**:A

Incorporating Query Expansion into Dynamic Match for Out-of-Vocabulary Word Detection

Zheng Yongjun, *Zhang Lianhai*

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

Abstract: To address the issue of out-of-vocabulary (OOV) word in keyword spotting, a method for incorporating query expansion into dynamic match is proposed. Query expansion and dynamic match are two different ways to compensate the high degree of uncertainty in OOV pronunciation. Considering the potential mutual complementarity between them, two fusion methods are presented. One is result fusion that performs a parallel OOV word detection with query expansion and dynamic match individually and then merges search results of the two systems. Another is confidence fusion which combines the minimum edit distance and the pronunciation score together as a hybrid confidence measure to implement OOV word detection and verification. Tests show that the second fusion method is more efficient and the figure of merit is superior to the baseline system which only uses dynamic match.

Key words: keyword spotting; query expansion; dynamic match; out-of-vocabulary word

引 言

关键词识别(Keyword recognition, KWR),亦称关键词检测(keyword spotting, KWS)[1],是指在语音数据中查找到所有可能出现的给定词的过程。语音关键词检测技术被看作是能有效处理口语和实现人机智能通信的解决方案之一[2]。目前 KWS 面临的一个主要挑战是集外词(Out-of-vocabulary, OOV)的检测。集外词是指那些不在系统字典里的词。出现集外词的原因是由于系统词表有一个固定的大小,不能覆盖全部的词汇,随着人类语言的变化发展,会产生更多的新词,这些词都是集外词,而它们也是人们关注的关键词。OOV 的检测性能相比集内词有一定差距,主要是因为 OOV 具有很高的发音不确定性和多样性,并且发音的不确定性很难通过声学和语言学模型来建模[3]。

解决 OOV 检测问题的常用方法是应用子词建模单元,例如:音素、音节、字形(Grapheme)、字形音素对(Graphone)和词片段(Word-fragment)等,首先将集外词转换为子词序列,然后将这些子词序列在先前创建的索引中检索。文献[4]根据声

基金项目:国家自然科学基金(61175017)资助项目;全军军事学研究生课题(2010JY0256-143)资助项目。

收稿日期:2013-12-19;**修订日期**:2014-01-09

学混淆度和语言模型得分将集外词扩展为集内词,弥补集外词的识别错误。文献[5]提出了动态匹配词格检索(Dynamic match lattice spotting, DMLS)方法,将基于音素 Lattice 的快速检测和动态序列匹配技术融合在一起,实现了快速而准确的关键词开集检测。文献[6]应用联合最大熵 N 元模型进行查询扩展,提升了语音文档检索的性能。文献[7]采用随机发音建模方法补偿集外词发音的不确定性,在索引中检测集外词所有可能的发音。文献[8]提出了前后缀查询扩展方法,并引入有穷自动机压缩检索空间,实现了高效的中文语音检索。文献[9-11]将多种语音索引系统融合在一起,应用不同子词建模单元间的互补性来提升集外词检测的性能。

动态匹配应用最小编辑距离(Minimum edit distance, MED)作为置信度,在检索时允许一定的误匹配,替换、插入和删除错误代价通过音素混淆矩阵得到,主要应用的是声学信息。而查询扩展通常是创建与集外词相关的多种发音,应用的是字形和发音之间的对应信息,没有应用声学信息。两者是在不同的层面补偿集外词发音的不确定性,理论上存在一定的互补性,基于此,本文将查询扩展和动态匹配融合在一起解决集外词的检测问题。首先利用 DMLS 方法搭建一个关键词检测系统,然后分析研究基于联合多元模型(Joint-multigram model, JMM)[12-13]的查询扩展和基于 MED 的动态匹配,最后将查询扩展和动态匹配融合在一起,采用了两种融合方法,一种是结果融合,另一种是置信度融合。实验结果表明二者的融合提升了系统性能。

1　基于 DMLS 的关键词检测系统

基于 DMLS 的关键词检测系统框架如图 1 所示。

索引阶段首先采用 BUT 的连续语音识别系统[14]生成音素 Lattice,音素 Lattice 提供了每个语音片段详细的音素表示形式,然后执行一个改进的维特比算法遍历 Lattice 来创建一个固定长度的音素序列数据库(Sequence database, SDB),作为后续检索操作的索引。后端检索阶段为研究的重点,本文主要研究集外词的检测。当一个集外词查询项提交给系统时,首先应用联合多元模型将集外词扩展为 n-best 发音的表示形式,并得到其发音的概率得分。其次,采用基于最小编辑距离的动态匹配确定与查询项发音近似匹配的音素序列。最后根据查询扩展和动态匹配的不同融合方法得到检测结果。

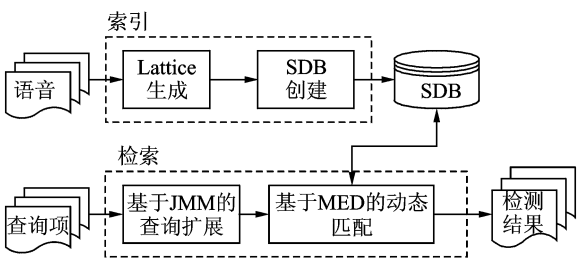

图 1　基于 DMLS 的关键词检测系统框架

Fig. 1　Architecture of keyword spotting system based on DMLS

2　融合查询扩展和动态匹配的改进

2.1　基于 JMM 的查询扩展

一个 LTS(Letter-to-sound)模型通常被定义为字形 $\boldsymbol{G}=(g_1,g_2,\cdots,g_L)$ 和发音 $\boldsymbol{Q}=(q_1,q_2,\cdots,q_R)$ 两个符号序列之间的随机映射。字形和发音是在同样的社会背景下发展起来的两个系统,两者之间具有紧密的联系,遵循不同的规则。字形和发音间的对应关系定义为字形和发音序列间的映射,其分量被称为字音对(Grapheme-phoneme pair)。最简单的映射为一个音素对应一个字形,如果字形和发音长度不同,可以插入空字符,此外多对多的映射也是合理的,例如图 2 给出了词"speaking"的字形和发音的对应关系。

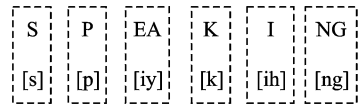

图 2　词"speaking"的字形和发音对应关系

Fig. 2　Grapheme-phoneme correspondence of the word "speaking"

JMM 的基本思想是对字形和发音的联合概率进行建模。多元 Multigram 是一个符号序列,长度可以为 0,1 或是更长。一个字音对包含一个字形多元和一个音素多元,因此也被称为联合多元。按照文献[12]的定义,一个联合多元 $\boldsymbol{u}=\{\tilde{g},\tilde{q}\}$ 被称为 graphone,其中 $\tilde{g}$ 为字形分量,$\tilde{q}$ 为音素分量。字形和发音的对应关系用 $\boldsymbol{U}$ 表示,实际为一个 graphone 序列,表示为

$$\boldsymbol{U}=(\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_H)=\begin{pmatrix}\tilde{g}_1,\tilde{g}_2,\cdots,\tilde{g}_H\\ \tilde{q}_1,\tilde{q}_2,\cdots,\tilde{q}_H\end{pmatrix} \quad (1)$$

式中 H 为 graphone 的长度,$\tilde{g}$ 和 $\tilde{q}$ 要满足如下约束

$$\begin{pmatrix}\widetilde{g}_1 \wedge \widetilde{g}_2 \wedge \cdots \wedge \widetilde{g}_H\\ \widetilde{q}_1 \wedge \widetilde{q}_2 \wedge \cdots \wedge \widetilde{q}_H\end{pmatrix}=\begin{pmatrix}g_1 g_2 g_3 \cdots g_L\\ q_1 q_2 q_3 \cdots q_R\end{pmatrix} \quad (2)$$

式中:符号∧表示连接;L 和 R 分别为字形和音素序列的长度;$\widetilde{g}$ 和 $\widetilde{q}$ 包含长度可变的符号。字形和发音的映射可以用 graphone 来描述。为了描述字形发音映射的随机属性,可以对 $\boldsymbol{U}$ 的概率分布进行建模,即联合多元模型,通常表示为

$$P(\boldsymbol{U})=P(\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_H) \quad (3)$$

在 JMM 中,字形 $\boldsymbol{G}$ 和发音 $\boldsymbol{Q}$ 的联合概率为所有可能 graphone 序列的概率总和,表示为

$$P(\boldsymbol{G},\boldsymbol{Q})=\sum_{\boldsymbol{U};\boldsymbol{G}(\boldsymbol{U})=\boldsymbol{G},\boldsymbol{Q}(\boldsymbol{U})=\boldsymbol{Q}}P(\boldsymbol{U})=\sum_{\boldsymbol{U};\boldsymbol{G}(\boldsymbol{U})=\boldsymbol{G},\boldsymbol{Q}(\boldsymbol{U})=\boldsymbol{Q}}P(\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_H) \quad (4)$$

式中 $\boldsymbol{G}(\boldsymbol{U})$和 $\boldsymbol{Q}(\boldsymbol{U})$分别为对应于 $\boldsymbol{U}$ 的字形和音素序列。那么发音预测公式为

$$\hat{Q}=\arg\max_{\boldsymbol{Q}}P(\boldsymbol{G},\boldsymbol{Q})=\arg\max_{\boldsymbol{Q}}\sum_{\boldsymbol{U};\boldsymbol{G}(\boldsymbol{U})=\boldsymbol{G},\boldsymbol{Q}(\boldsymbol{U})=\boldsymbol{Q}}P(\boldsymbol{U}) \quad (5)$$

式中 $P(\boldsymbol{U})$可以应用标准的 n 元语言模型建模,得到

$$P(\boldsymbol{U})=\prod_{j=1}^{H}P(\boldsymbol{u}_j \mid \boldsymbol{u}_{j-1},\cdots,\boldsymbol{u}_{j-n+1}) \quad (6)$$

本文应用工具包 Sequitur G2P[15] 训练 JMM 模型,n-best 发音预测解码算法详见文献[12]。应用 JMM 模型可以将集外词查询项 term 扩展为其 n-best 发音的表示形式$(\boldsymbol{Q}_1,P(\boldsymbol{Q}_1\mid\boldsymbol{G})),\cdots,(\boldsymbol{Q}_n,P(\boldsymbol{Q}_n\mid\boldsymbol{G}))$,$\boldsymbol{Q}_i$ 为一个发音,$P(\boldsymbol{Q}_i\mid\boldsymbol{G})$为其相应的发音得分,代表这一发音的概率。以词“bungalow”的查询扩展为例,如表 1 所示。每一个发音 $\boldsymbol{Q}_i$ 的置信度可以用其发音得分的对数表示,如式(7)所示。在集外词检测时,如果在索引中同一时段内检测到查询项 term 的多个发音时,将检测结果合并并分配最大的发音得分置信度,如式(8)所示。

$$C_{\text{pron}}(\boldsymbol{Q}_i)=\log(P(\boldsymbol{Q}_i\mid\boldsymbol{G})) \quad (7)$$

$$C_{\text{JMM}}(\text{term})=\max_i C_{\text{pron}}(\boldsymbol{Q}_i) \quad (8)$$

表 1 词“bungalow”的查询扩展

Table 1 Query expansion of the word "bungalow"

发音	关键词							得分
参考发音	b	ah	ng	g	ah	l	ow	—
	b	ah	ng	g	ah	l	ow	0.922
	b	ah	ng	g	ah	l	aw	0.030
6-best	b	ah	ng	g	aa	l	ow	0.012
发音	b	ah	n	g	ah	l	ow	0.008
	b	ah	ng	g	ey	l	ow	0.005
	b	uw	ng	g	ah	l	ow	0.004

2.2 基于 MED 的动态匹配

在语音识别中,经常会出现替换、插入和删除错误(见图 3),并且集外词出现识别错误的概率更高,这严重影响关键词检测的性能。因此在检索中采用动态匹配,应用最小编辑距离作为置信度,允许一定的误匹配来补偿识别错误。

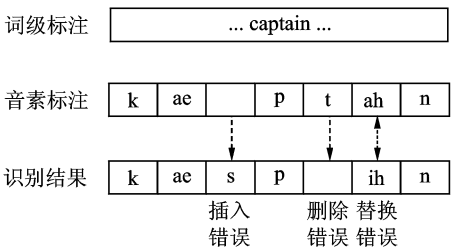

图 3 语音识别常见错误

Fig. 3 Common errors in speech recognition

最小编辑距离又称为 Levenshtein 距离,主要用于度量将一个字符串转换为另外一个字符串所付出的最小代价。这种转换主要包含 4 种编辑操作:匹配、替换、插入和删除,通常每种编辑操作都会有相应的代价。本文应用 MED 检测和查询项 term 的发音 $\boldsymbol{Q}$(目标音素序列)近似匹配的索引音素序列。MED 计算的核心思想是应用一个代价矩阵来累加转换代价,执行一个递推过程来更新代价矩阵的各个元素,从而确定整体的最小转换代价。定义 $\boldsymbol{\Phi}=(\phi_1,\phi_2,\cdots,\phi_N)$ 为索引音素序列,$\boldsymbol{Q}=(q_1,q_2,\cdots,q_M)$为目标音素序列,$\boldsymbol{\Omega}_{N+1\times M+1}$为 $N+1\times M+1$ 维的代价矩阵,$C_s(\phi_i,q_j)$,$C_i(\phi_i)$和 $C_d(q_j)$分别为替换、插入和删除代价 $\Omega_{i,j}$为代价矩阵 $\boldsymbol{\Omega}$ 中的元素,表示将子序列$(\phi)_1^i$ 变换为$(q)_1^j$ 的最小代价。具体流程如下:

(1)初始化一个 $N+1\times M+1$ 维的代价矩阵 $\boldsymbol{\Omega}$,矩阵的第一个元素 $\Omega_{0,0}=0$;

(2)初始化代价矩阵的第一行元素

$$\Omega_{0,j}=\sum_1^j C_d(q_j) \qquad 0<j\leqslant M \quad (9)$$

(3)初始化代价矩阵的第一列元素

$$\Omega_{i,0}=\sum_1^i C_i(\phi_i) \qquad 0<i\leqslant N \quad (10)$$

(4)从左到右从上到下依次更新代价矩阵的元素

$$\Omega_{i,j}=\text{Min}\begin{pmatrix}\Omega_{i,j-1}+C_d(q_j)\\ \Omega_{i-1,j}+C_i(\phi_i)\\ \Omega_{i-1,j-1}+C_s(\phi_i,q_j)\end{pmatrix} \qquad 0<i\leqslant N,0<j\leqslant M \quad (11)$$

$\Omega_{N,M}$即为将 $\boldsymbol{\Phi}$ 转换为 $\boldsymbol{Q}$ 的最小代价,通常情况下 $N \geqslant M$,所以还要执行一个回溯算法寻找最优路径,并在最优路径上确定和 $\boldsymbol{Q}$ 最为近似的音素子序列 $\boldsymbol{\Phi}'$,得到时间边界信息和最小编辑距离 $\Delta(\boldsymbol{\Phi}',\boldsymbol{Q})$。在 DMLS 中,某一检测结果的置信度得分被简单定义为索引音素子序列 $\boldsymbol{\Phi}'$ 和目标音素序列 $\boldsymbol{Q}$ 之间距离的负值,如式(12)所示。如果 $C_{\mathrm{MED}}(\boldsymbol{\Phi}',\boldsymbol{Q})$在设定的阈值范围之内,那么 $\boldsymbol{\Phi}'$ 就为检测结果。

$$C_{\mathrm{MED}}(\boldsymbol{\Phi}',\boldsymbol{Q}) = -\Delta(\boldsymbol{\Phi}',\boldsymbol{Q}) \tag{12}$$

在 MED 的原始定义中,替换、插入和删除的代价通常为 1,不能完全反映各个音素之间识别错误的规律和模糊发音现象。因此可以通过观察音素识别器实际产生的音素错误训练得到改进的替换、插入和删除错误代价[16]。HTK 工具包中的 HResults 被用于对齐训练集音素识别结果和参考的音素标注,生成一个音素混淆矩阵。替换、插入和删除错误代价可以通过最大似然估计从音素混淆矩阵中训练得到。应用音素混淆度加权的 MED 可以更好地补偿音素识别错误,从而改善关键词检测的准确性。

2.3　融合查询扩展和动态匹配

由上文可知,查询扩展和动态匹配是在不同的层面补偿集外词发音的不确定性,理论上存在一定的互补性,可以将二者融合在一起进行集外词的检测。本文研究了两种融合方法,第一种是结果融合(如图 4 所示)。分别应用查询扩展和动态匹配并行的检测集外词。查询扩展:应用 JMM 将集外词查询项 term 扩展为 n-best 发音 $\boldsymbol{Q}_i$,$1 \leqslant i \leqslant n$,然后将这些发音在索引中进行精确匹配,得到检测结果。动态匹配:同样是应用 JMM 得到 term 的 1-best发音 $\boldsymbol{Q}_{1\text{-best}}$,然后在索引中采用动态匹配,检索和 $\boldsymbol{Q}_{1\text{-best}}$近似的结果。最后将两个系统在索引中的同一时间段内出现的检测结果合并,并分配最大的置信度得分(如式(13)),同时保留不同的检测结果。

$$\boldsymbol{C}_{\mathrm{MED+JMM}}(\mathrm{term}) = \max(C_{\mathrm{MED}}(\boldsymbol{\Phi}',\boldsymbol{Q}_{1\text{-best}}), C_{\mathrm{JMM}}(\mathrm{term})) \tag{13}$$

另外一种方法是置信度融合,最小编辑距离衡量的是查询项发音(目标音素序列)和索引音素序列间的相似度,而发音得分描述的是发音和字形之间的对应关系,两种置信度都是和发音相关的,且在同一个数量级上,可以将 MED 和发音得分融合构成混合置信度进行集外词的检出和确认,如式(14)所示。

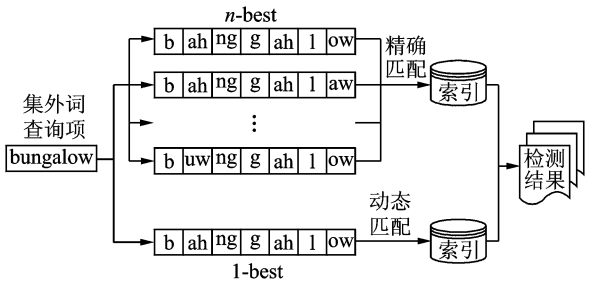

图 4　查询扩展和动态匹配的结果融合

Fig. 4　Result fusion between query expansion and dynamic match

$$C_{\mathrm{fusion}}(\boldsymbol{\Phi}',\boldsymbol{Q}_i) = \eta C_{\mathrm{MED}}(\boldsymbol{\Phi}',\boldsymbol{Q}_i) + (1-\eta) C_{\mathrm{pron}}(\boldsymbol{Q}_i) \tag{14}$$

$$C_{\mathrm{fusion}}(\boldsymbol{\Phi}',\mathrm{term}) = \max_i C_{\mathrm{fusion}}(\boldsymbol{\Phi}',\boldsymbol{Q}_i) \quad 1 < i \leqslant n \tag{15}$$

式中 η 为加权因子,平衡 MED 和发音得分两种置信度的贡献度。在实际检测中,同一个查询项 term 的多个发音可能会出现在索引中的同一时间段内,需要合并检测结果并分配最大的置信度得分,如式(15)所示。实验表明置信度融合的方法更好,有效提升了系统的性能。

2.4　算法步骤

(1)采用连续语音识别系统生成音素 Lattice;

(2)执行一个改进的维特比算法遍历 Lattice 来创建索引;

(3)应用发音字典 CMUdict 训练 JMM 模型;

(4)应用 JMM 模型将集外词查询项扩展为其 n-best 音素发音;

(5)应用两种不同的融合方法在索引中检索关键词,并根据置信度阈值输出检测结果。

3　实验结果及分析

3.1　实验配置

本文实验采用 TIMIT 语料库,主要分为 TRAIN 和 TEST 两个文件集合。本文实验选择 TRAIN 中 3 696 个语句作为训练集,选择 TEST 中 1 344 个语句作为测试集,未采用其中适合于说话人实验的 SA1 和 SA2 中的语句。TIMIT 语料库中共含有 61 个音素单元,其划分较为精细,按照 BUT 的划分标准,将 TIMIT 中 61 个音素映射为 39 个音素,如将塞音的成阻(Closure)和除阻(Burst)部分合并(bcl b→b)。实验选取的集外词规模为 100 个,各关键词在测试集 TEST 中出现

的次数总共为 344 次。训练 JMM 模型应用的是卡耐基梅隆大学的英文发音字典 CMUdict，该字典共包含 125 000 个英文单词，训练时剔除了 1 832个和实验选取的集外词相关的单词。

3.2 评价标准

本文采用接收机工作特性(Receiver operating characteristics，ROC)曲线和品质因数(Figure of merit，FOM)[17]作为系统性能的评价指标。ROC 曲线定义为不同的置信度阈值下，系统的召回率 P_{Recall} 随虚警率 P_{FA} 的变化趋势，反映了系统的综合性能。召回率 P_{Recall} 为正确的关键词检测结果数量 N_{correct} 占实际出现的关键词数量 N_{true} 的百分比。虚警率 P_{FA} 定义为虚警个数 N_{FA} 被分母归一化后的结果，H 为语音文档长度，S 为关键词词表大小，如式(16，17)所示。FOM 定义为虚警率在 0～10 范围内的平均召回率，如式(18)所示。

$$P_{\text{Recall}}=\frac{N_{\text{correct}}}{N_{\text{true}}} \tag{16}$$

$$P_{\text{FA}}=\frac{N_{\text{FA}}}{H\cdot S} \tag{17}$$

$$\text{FOM}=\frac{1}{10}\int_{P_{\text{FA}}=0}^{P_{\text{FA}}=10}P_{\text{Recall}}\,\text{d}(P_{\text{FA}}) \tag{18}$$

3.3 系统性能比较

将基于 JMM 查询扩展的检测方法记为 C_{JMM}，基于动态匹配的检测方法记为 C_{MED}，第一种融合查询扩展和动态匹配的检测方法记为 $C_{\text{MED+JMM}}$，第二种融合方法记为 C_{fusion}。表 2 对比了不同系统集外词检测的性能，图 5 给出了相应的 ROC 曲线。从图中可以看出 C_{MED} 方法优于 C_{JMM} 方法，主要是由于动态匹配在检索中综合考虑了音素识别的错误规律，能够更好地补偿集外词的不确定性。同时，两种方法又是在不同的层面补偿集外词发音的不确定性，理论上存在一定的互补性，实验结果也验证了这一点。第一种融合方法相比单一的动态匹配，FOM 相对提升了 3.9%，说明直接将两种方法的检测结果融合具有一定的互补性，但性能提升有限，需要进一步进行优化融合方法。

在第二种融合方法中，查询扩展的阶数 n 和加权因子 η 共同决定系统的最终性能。表 3 给出了 C_{fusion} 方法在不同扩展阶数 n 和加权因子 η 下系统的性能。当查询扩展为 2-best 发音，加权因子 η=0.7 时，FOM 相对提升了 19.8%，具有最优的系统性能。这说明在优化了扩展阶数和置信度得分贡献度后，第二种融合方法效果更好。另外，分析表 3 的实验结果可知在动态匹配中已经允许一定的误匹配存在，如果查询扩展阶数较大，虚警率将急剧增加，影响整体性能，当扩展为 3-best 发音时，FOM 下降已经非常明显。

表 2 不同系统集外词检测性能的比较

Table 2 Comparison of OOV detection performance in different systems

系统描述	FOM
C_{JMM}	0.227 8
C_{MED}	0.354 7
$C_{\text{MED+JMM}}$	0.368 6
C_{fusion}(2-best，η=0.7)	0.424 8

表 3 C_{fusion} 方法在不同参数下的 FOM

Table 3 FOM of C_{fusion} method using different parameters

η	n-best	
	2-best	3-best
0.9	0.408 8	0.286 3
0.8	0.422 1	0.284 4
0.7	0.424 8	0.269 6
0.6	0.423 7	0.247 1
0.5	0.405 2	0.213 1

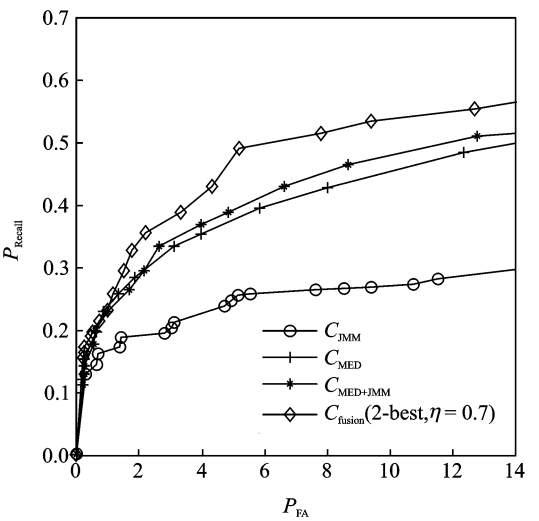

图 5 不同集外词检测系统的 ROC 曲线

Fig. 5 ROC curves of various OOV detection systems

4 结束语

针对关键词检测中集外词检测性能较低的问题，本文提出了一种改进的集外词检测方法，将基于联合多元模型的查询扩展和基于最小编辑距离的动态匹配融合在一起。本文研究了两种融合方

法：第一种方法是直接将两者的检测结果进行融合，性能有一定提升；第二种方法通过引入一个加权因子 η 来优化平衡最小编辑距离和发音得分置信度的贡献度。实验结果表明，在扩展为 2-best 发音，加权因子 $\eta=0.7$ 时，查询扩展和动态匹配具有最优的互补性，FOM 相对提高了 19.8%。下一步的工作重点是研究混合索引问题，提升系统的实用性。

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] 孙成立. 语音关键词识别技术的研究[D]. 北京：北京邮电大学，2008：1-2.
Sun Chengli. A study of speech keyword recognition technology [D]. Beijing: Beijing University of Posts and Telecommunications, 2008:1-2.

[3] Wang Dong. Out-of-vocabulary spoken term detection[D]. Edinburgh : School of Informatics, University of Edinburgh, 2010:9-13.

[4] Logan B, Thong J M V. Confusion-based query expansion for OOV words in spoken document retrieval [C]//The 7th International Conference on Spoken Language Processing. Colorado, USA: ISCA, 2002: 1997-2000.

[5] Thambiratmann K, Sridharan S. Rapid yet accurate speech indexing using dynamic match lattice spotting [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007,15(1):346-357.

[6] Mamou J, Ramabhadran B. Phonetic query expansion for spoken document retrieval[C]//The 9th Annual Conference of the International Speech Communication Association. Brisbane, Australia: ISCA, 2008:2106-2109.

[7] Wang Dong, King S, Frankel J. Stochastic pronunciation modeling for out-of-vocabulary spoken term detection[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011,19(4):688-698.

[8] 李伟，吴及，吕萍. 基于查询扩展的中文语音高效检索[J]. 模式识别与人工智能，2011，24(4)：561-566.
Li Wei, Wu Ji, Lü Ping. Query expansion based high performance Chinese voice retrieval[J]. Pattern Recognition and Artificial Intelligence, 2011,24(4):561-566.

[9] Qin Long, Sun Ming, Rudnicky A. System combination for out-of-vocabulary word detection[C]// IEEE International Conference on Acoustic, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012:4817-4820.

[10] Xu Yong, Guo Wu, Dai Lirong. A hybrid fragment/syllable-based system for improved OOV term detection[C]// The 8th International Symposium on Chinese Spoken Language Processing. Hong Kong, China:[s. n.],2012:378-382.

[11] Kanda N, Itoyama K, Okuno H G. Multiple index combination for Japanese spoken term detection with optimum index selection based on OOV-region classifier[C]// IEEE International Conference on Acoustic, Speech and Signal Processing. Vancouver, Canada: IEEE,2013:8540-8544.

[12] Bisani M, Ney H. Joint-sequence models for grapheme-to-phoneme conversion[J]. Speech Communication, 2008,50(5):434-451.

[13] Jouvet D, Fohr D, Illina I. Evaluating grapheme-to-phoneme converters in automatic speech recognition context[C]// IEEE International Conference on Acoustic, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012:4821-4824.

[14] Grezl F, Karafiat M. Integrating recent MLP feature extraction techniques into TRAP architecture[C]// The 12th Annual Conference of the International Speech Communication Association. Florence, Italy: ISCA, 2011:1229-1232.

[15] Hahn S, Lehnen P, Wiesler S, et al. Improving LVCSR with hidden conditional random fields for grapheme-to-phoneme conversion[C]//The 14th Annual Conference of the International Speech Communication Association. Lyon, France: ISCA, 2013: 495-499.

[16] Wallace R. Fast and accurate phonetic spoken term detection [D]. Queensland: Queensland University of Technology, 2010:51-90.

[17] Wallace R, Baker B, Vogt R, et al. Discriminative optimization of the figure of merit for phonetic spoken term detection[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(6): 1677-1687.

作者简介：郑永军(1984-)，男，硕士研究生，研究方向：语音识别、模式识别，E-mail：banjiu123cool@163.com；张连海(1971-)，男，副教授，研究方向：语音信号处理、模式识别。