基于格的汉语自然对话语音索引方法研究

孟 莎[1] 余 鹏[2] 刘 加[1]

摘 要 对汉语自然对话语音索引问题进行了研究. 比较了不同单元格的识别和检索性能, 提出不同单元格的转换方法、格间的融合方法以及格内节点与边的合并方法. 格转换实现了识别单元和索引单元的分离, 词格转换得到的无调音节格将品质因数 (Figure of merit, FOM) 从基线系统的 69.2 % 提高到 73.7 %; 格间融合综合利用多个格的信息, 将 FOM 进一步提高到 78.6 %; 格内合并对格进行了有效的压缩, 使其可应用于海量语音检索.

关键词 语音检索, 语音索引, 后验概率格, 索引单元
DOI 10.3724/SP.J.1004.2010.00215

Lattice-based Indexing for Spontaneous Mandarin Speech

MENG Sha[1] YU Peng[2] LIU Jia[1]

Abstract We examine the task of spoken term detection in Chinese spontaneous speech with a lattice-based approach. We compare lattices generated with different units and lattices converted from one unit to another. We find that the best system is with toneless-syllable lattices converted from word lattices whose figure of merit (FOM) is 73.7 % from the baseline 69.2 %. By combining lattices from multiple systems into a single lattice and fully exploiting the redundant information in the combined lattice with a time-based node/arc merging, we achieve the result of a compact lattice index with the accuracy improved up to 79.2 %.

Key words Speech retrieval, speech indexing, posterior lattice, indexing units

随着信息技术、多媒体技术和互联网技术的发展, 大量的语音数据迅速积累, 如何对其进行有效索引和快速查找是亟待解决的问题. 基于内容的语音检索技术致力于从大规模语音数据中快速找到与查询词相关的段落, 是当前的研究热点之一. 图 1 所示是一个通用的语音检索系统. 与文本检索方法类似, 语音检索分为索引和查找两个阶段: 图 1 中实线部分为索引阶段, 通常首先通过语音识别技术对语音数据进行处理并建立索引文件, 整个索引过程在后台完成, 对每个文件只做单次处理; 虚线部分为查询阶段, 在用户每次查询时进行 (多次), 需要很高的实时性.

对语音检索系统, 长期存在两个方面的讨论: 基于听写机的检索系统与基于格的检索系统的比较; 基于不同识别单元和索引单元的检索系统 (对英文而言, 通常指词与音素) 的比较.

收稿日期 2008-11-14 录用日期 2009-04-08
Manuscript received November 14, 2008; accepted April 8, 2009
国家高技术研究发展计划 (863 计划) (2006AA010101, 2007AA04Z223), 国家自然科学基金委员会与微软亚洲研究院联合资助项目 (60776800) 资助
Supported by National High Technology Research and Development Program of China (863 Program) (2006AA010101, 2007AA04Z223) and National Natural Science Foundation of China and Microsoft Research Asia (60776800)
1. 清华大学电子工程系清华信息科学与技术国家实验室 (筹) 北京 100084 2. 微软亚洲研究院 北京 100190
1. Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Electronic Engineering, Tsinghua University, Beijing 100084 2. Microsoft Research Asia, Beijing 100190

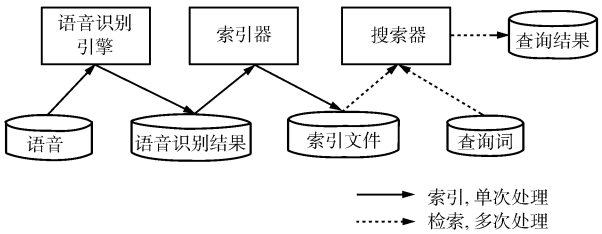


图 1 语音检索系统

Fig. 1 Spoken document retrieval system scheme

早期的语音检索系统大多采用直接将语音识别引擎和信息检索引擎结合的方法, 即用语音听写机的结果进行索引, 供用户查询. 基于听写机的系统中, 语音识别和信息检索是相互独立的模块, 后者可以直接利用文本信息检索领域已有的方法和系统, 特别是能够处理海量数据的网络搜索引擎. 美国国家标准技术研究所 (National Institute of Standards and Technology, NIST) 于 1997 年 ∼ 2000 年进行的语音检索评测即采用这种方法, 在广播语音的识别结果上得到了和在人工标注文本上相当的检索性能, 因此认为这是 "一个已经解决的问题"[1]. 但后续的研究表明, 这种方法并不适用于检索自然对话语音. 首先广播语音质量较好, 识别准确率高, 词错误率 (Word error rate, WER) 一般在 20 % 以内, 而自然对话语音的准确率较低, WER 往往高于 30 %. 另一方面测试集段落较长且有重复, 降低了语音识别错误对检索任务的影响.

语音识别引擎在生成最大似然识别结果的同时, 可提供似然度相对较小的其他候选识别结果, 以格 (Lattice) 的形式保存. 最大似然识别结果有比较高的错误率, 格中往往包含了大量正确的补充信息. 在小规模测试集上的研究表明, 对语音识别结果格进行索引可以极大地提高系统的查全率[2−3], NIST 于 2006 年开始的语音查询词检测 (Spoken term detection, STD) 评测系统多采用基于格的方法. 但对于构建实用系统, 尤其是海量语音数据的检索系统而言, 格的引入会占用极大的索引空间, 如何对格索引进行压缩是一个重要问题[4].

识别单元和索引单元的选择对系统检索性能有着极大的影响. 通常情况下, 识别和索引采用相同的单元. 在英文语音检索领域, 对选择词或音素作为识别索引单元有着广泛的讨论. 基于词的系统[1, 4−5] 因不能解决集外词问题, 即无法识别和检索出没有包含在字典和语言模型中的词, 严重影响了查全率; 而基于音素[3, 6−7] 的系统虽不存在集外词问题, 但由于语言模型较弱, 识别性能和查准率大大降低. 一些研究[8−9] 将两类系统进行融合得到了较好的效果. 汉语与英语有很大不同, 常见的单元有词、字和音节等, 已有系统大多基于音节进行[10−12], 文本检索领域对中文中词和字的索引性能进行了研究[13], 我们在早期的工作中对词和音节的识别和索引性能进行了对比[14].

本文基于格建立语音索引, 将格进行后验概率表示, 对不同单元 — 词、字、音节、无调音节的识别和索引性能进行比较. 基于后验概率格的特性, 我们提出不同单元的格的转换方法、多个格间的融合方法以及格内节点和边的合并方法. 通过不同单元格的转换, 在识别和索引中可采用不同的单元; 对不同来源的格的融合极大地提高了整体系统的检索性能; 格内部节点和边的合并对格进行了有效的压缩, 将规模控制在可接受的范围, 为大规模的语音数据索引提供了必要条件.

1 基于格的语音检索系统

1.1 格

识别过程中得到的格由节点和边组成, 每一个节点 (Node) 表示一个时间点, 每一条边 (Arc) 表示一个识别单元假设 (词或子词), 边将不同时刻的节点连接起来, 形成一个有向非循环图. 一个格的信息可以表示为: $L = (N, A, n_{\text{start}}, n_{\text{end}})$, 其中 N, A 分别表格中节点与边的集合, n_{start}, n_{end} 表示格的开始节点和结束节点. 每一节点包含对应的时间信息 $t(n)$ 和上下文信息; 每一条边 a 可表示为一个四元组 $(S[a], E[a], I[a], w[a])$, $S[a]$ 和 $E[a]$ 分别表示这条边的起始节点和结束节点, $I[a]$ 表示识别单元假设, $w[a]$ 表示识别单元假设的似然度, 典型的表示方法是语音特征与声学模型匹配的似然度 $P_{ac}(a)$ 和带权重因子 λ 的上下文信息与语言模型匹配的似然度 $P_{LM}(a)$ 的乘积, 如式 (1) 所示.

$$w[a] = P_{ac}(a) \cdot P_{LM}^{\lambda}(a) \tag{1}$$

1.2 格的后验概率表示

格的似然度表示方法局部考虑当前边的声学特征与声学模型匹配的似然程度和上下文环境与语言模型匹配的似然程度, 在全局上不具有可比性; 格的后验概率表示解决了这个问题.

将语音识别结果的一条路径记为: $\pi = (a_1, a_2, \cdots, a_K)$, $a_1, a_2, \cdots, a_K$ 为路径经过的所有边, 则该路径的起始节点 $S[\pi] = S[a_1]$, 终止节点 $S[\pi] = S[a_K]$, 识别结果为 $I[\pi] = (I[a_1], I[a_2], \cdots, I[a_K])$, 似然度为 $w[\pi] = \prod_{k=1}^{K} w[a_k]$. 格中节点的后验概率 $P_{\text{node}}[n]$ 和边的后验概率 $P_{\text{arc}}[a]$ 分别表示为

$$P_{\text{node}}[n] = \frac{\alpha_n \cdot \beta_n}{\alpha_{n_{\text{end}}}} \tag{2}$$

$$P_{\text{arc}}[a] = \frac{\alpha_{S[a]} \cdot w[a] \cdot \beta_{E[a]}}{\alpha_{n_{\text{end}}}} \tag{3}$$

其中

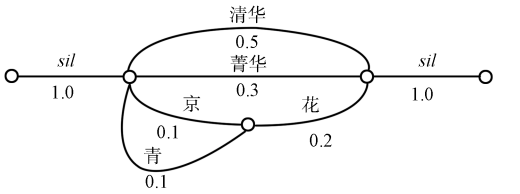
$$\alpha_n = \sum_{\pi: S[\pi]=n_{\text{start}} \wedge E[\pi]=n} w(\pi) \tag{4}$$

$$\beta_n = \sum_{\pi: S[\pi]=n \wedge E[\pi]=n_{\text{end}}} w(\pi) \tag{5}$$

α 和 β 根据前后向方法计算可得[15]. 计算格中所有节点和边的后验概率, 则可得到格的后验概率表示. 后验概率词格中节点为二元组 $(t[n], P_{\text{node}}[n])$, 边仍为四元组 $(S[a], E(a), I[a], P_{\text{arc}}[a])$. 词组 Q (包含多条边) 的后验概率可由下式计算[4]:

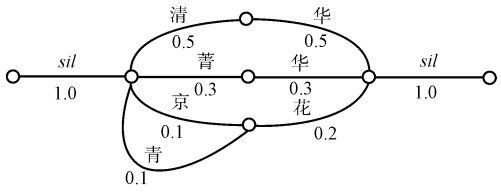
$$\begin{aligned} P(*, t_s, Q, t_e, *|O) = \\ \sum_{\substack{\pi=(a_1,a_2,\cdots,a_K): \\ t[S[\pi]]=t_s \wedge t[E[\pi]]=t_e \wedge I[\pi]=Q}} \frac{\alpha_{S[a_1]} \prod_{i=1}^{K} w[a_i] \cdot \beta_{E[a_K]}}{\alpha_{n_{\text{end}}}} = \\ \sum_{\substack{\pi=(a_1,a_2,\cdots,a_K): \\ t[S[\pi]]=t_s \wedge t[E[\pi]]=t_e \wedge I[\pi]=Q}} \frac{P_{\text{arc}}[a_1] \times \cdots \times P_{\text{arc}}[a_K]}{P_{\text{node}}S[a_2] \times \cdots \times P_{\text{node}}S[a_K]} \end{aligned} \tag{6}$$

后验概率表示的词格有两个重要特性: 1) 对词格中的每一个节点, 所有进入节点的边的后验概率和等于所有从节点发出的边的后验概率和; 2) 对每一个时间点, 所有经过这个时间点的边的概率和为 1, 如图 2 (a) 所示.



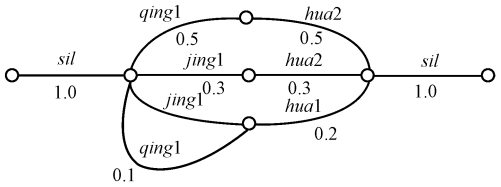
(a) 词格

(a) Word lattice



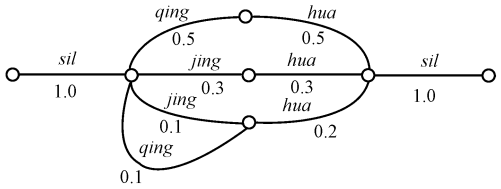
(b) 字格

(b) Character lattice



(c) 音节格

(c) Syllable lattice



(d) 无调音节格

(d) Toneless syllable lattice

图 2　后验概率格

Fig. 2　Posterior lattices

2　格的产生和转换

2.1　语音识别与格的产生

我们采用当前主流的大词汇量连续语音识别器 (Large vocabulary continuous speech recognition, LVCSR), 分别选择词、字、音节和无调音节作为识别单元对语音数据进行解码. 当选用词作为识别单元时, 根据所用词表, 首先对语言模型训练语料进行分词, 而后训练出基于词的 N 元 (n-gram) 语言模型. 当选用字、音节和无调音节时, 相同的语料分别用于训练出基于字、音节和无调音节的 N 元语言模型. 在识别过程中采用与语言模型对应的发音字典, 即可分别生成词格、字格、音节格和无调音节格[16].

2.2　不同单元格的转换

常见 LVCSR 系统往往采用词作为识别单元. 在英文中, 以词作为识别单元的最大问题是集外词问题; 对汉语语音检索而言, 在识别过程中可以通过对所有汉字建模, 集外词可看作是汉字或集内词的合成词, 也有被正确识别和检测的可能. 但汉语存在分词的多样性问题, 当查询词与识别结果的分词不匹配时, 不能得到命中, 例如若识别词格结果为 “清华大学”, 当关键词查询为 “清华” 时不能得到命中. 因此, 直接利用 LVCSR 的识别结果词格作为索引不能最大化利用识别结果信息. 一个有效的解决方法是将词格分解成子词 (字, 音节等) 格. 设词格中的一条边 $I[a] = W = (c_1, c_2, \cdots, c_N)$, 其中 W 是识别单元词, $c_1, c_2, \cdots, c_N$ 为组成 W 的字序列, 则词格到字格的转换方法如下.

对词格中所有的边 a:

1) 新建节点 $n_1, \cdots, n_{N-1}$, 令

$$\begin{aligned} t[n_i] &= \frac{(i \times t[E[a]] + (N-i) \times t[S[a]])}{N} \\ P_{\text{node}}[n_i] &= P_{\text{arc}}[a] \end{aligned}$$

2) 新建边 $a_1, \cdots, a_N$, 令

$$S[a_i] = \begin{cases} S[a]\colon\ i = 1 \\ n_{i-1}\colon\ i > 1 \end{cases};\ E[a_i] = \begin{cases} n_i\ \ \colon\ i < N \\ E[a]\colon\ i = N \end{cases}$$

$$P_{\text{arc}}[a_i] = P_{\text{arc}}[a]; \qquad I[a_i] = c_i$$

3) 删除 a

汉语的另一重要特点是存在大量的同音字、同音词. 识别过程中常存在声学识别正确, 但因为语言信息的不充分, 识成相同发音的其他字或词. 考虑到这个因素, 我们将词格转化为音节格. 语调信息也是易混淆信息之一, 因此又考虑将有调音节格转换为无调音节格. 图 2 (a) 中所示词格可分别转换成图 2 (b) ∼ 2(d) 所示的字格、音节格和无调音节格. 通过不同类型格的转换, 索引单元可以采用与识别不同的单元.

3　格间融合

通过不同的识别单元选择和格的转换, 有不同的途径得到相同单元或不同单元的格. 考虑将不同来源的格进行融合, 以综合利用多个格所包含的信息互相补充, 可以采用后向融合和前向融合两种方法.

3.1　后向融合

后向融合是对检索结果的融合, 在搜索时进行. 对不同来源的多个格分别进行索引, 用户搜索时, 多个格索引同时返回结果, 对这些结果进行融合[16]. 设 $L_1, L_2, \cdots, L_n$ 为不同来源的格, $P(*, t_s, Q, t_e, *|L_i)$ 为查询词 Q 在 L_i 匹配得到的

后验概率, 则融合后的后验概率为

$$P^{COMB}(*,t_s,Q,t_e,*)=\sum_{i=1}^{n}\gamma_i\cdot P(*,t_s,Q,t_e,*|L_i) \quad (7)$$

其中, $\sum_{i=1}^{n}\gamma_i=1$. 实验证明, 权重因子 γ 的选择影响很小, 我们选取 $\gamma_i=1/n$.

3.2 前向融合

前向融合是对格的融合, 在建立索引时进行. 将不同词格的起始节点与起始节点之间, 结束节点与结束节点之间对应合并, 构成更丰富的格索引, 融合得到的格中, 边与节点的后验概率由下式计算得到:

$$P'_{\rm arc}[a]=\gamma_i\cdot P_{\rm arc}[a] \quad (8)$$

$$P'_{\rm node}[n]=\gamma_i\cdot P_{\rm node}[n] \quad (9)$$

其中 γ_i 的取值与后向融合相同. 图 3 (a) 与 3 (b) 所示的两个后验概率音节格融合得到的结果如图 3 (c) 所示. 前向融合和后向融合有相同的效用 — 相同的性能提升和整体格规模, 但前向融合将所有来源的格统一到一个索引之中, 只需搜索一次. 下一节将说明前向融合得到的格更有利于格规模的压缩. 本文的实验部分采用前向融合方法.

4 格内合并

对于基于格的语音检索系统, 尤其是进行融合之后, 如何进行压缩是将其应用于大规模检索系统的关键问题. 首先我们对格中的冗余成分进行分析, 通过去除冗余来降低格的规模, 这样可以不损害系统性能. 多个来源的格中包含格内和格间两类冗余.

1) 格内冗余: 语音识别生成的格中, 由于不同的上下文信息, 往往多条边对应了相同的识别单元假设 (词、字、音节和无调音节等) 和相同的时间信息, 这在语音检索任务中是不必要的. 例如, 图 2(c) 所示音节格中, 在一个时间段, 有 2 条标识为 “qing1” 的边和 2 条标识为 “jing1” 的边.

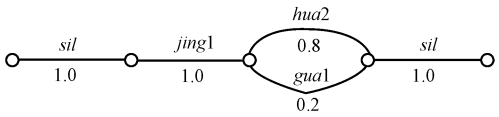
2) 格间冗余: 不同来源的具有相同单元的格中, 往往存在大量具有相同识别单元假设和相同时间信息的边. 这些边的重复存在也是冗余的. 如图 3 (a) 和图 3 (b) 中在同一时间段有 2 条标识为 “jing1” 的边.

格内冗余可通过格内部节点和边的合并来消除.

1) 节点合并: 若存在一组节点 $\{n_i, i=1,2,\cdots,N\}$, 满足 $t[n_i]=t$:

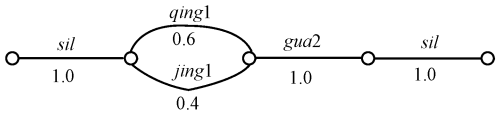
a) 新建节点 n, 令 $t[n]=t$, $P_{\rm node}[n]=\sum_{i=1}^{N}P_{\rm node}[n_i]$;

b) $\forall a\in\{a|S[a]\in\{n_i,\ i=1,2,\cdots,N\}\}$, 令 $S[a]=n$;



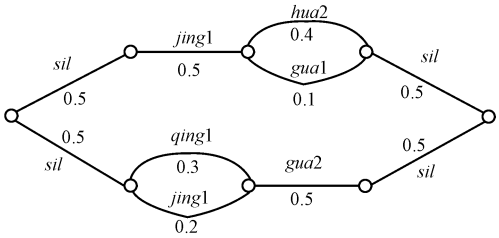
(a) 音节格 1

(a) Syllable lattice 1



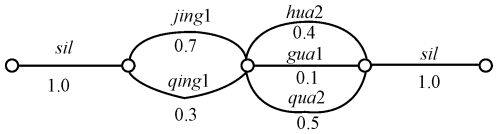
(b) 音节格 2

(b) Syllable lattice 2



(c) 格间融合

(c) Lattice fusion



(d) 格内合并

(d) Lattice merging

图 3 格间融合与格内合并

Fig. 3 Lattice fusion and merging

c) $\forall a\in\{a|E[a]\in\{n_i,\ i=1,2,\cdots,N\}\}$, 令 $E[a]=n$;

d) 删除 n_i, $i=1,2,\cdots,N$.

2) 边合并: 若存在一组边 a_j, $j=1,2,\cdots,K$, 满足 $I[a_j]=I$, $S[a_j]=S$, $E[a_j]=E$:

a) 新建边 a, 令 $I[a]=I$, $S[a]=S$, $E[a]=E$, $P_{\rm arc}(a)=\sum_{j=1}^{K}P_{\rm arc}(a_k)$;

b) 删除 a_j, $j=1,2,\cdots,K$.

通过格间的前向融合, 多个格融合在一起成为一个格, 格间冗余转化为格内冗余, 通过上述方法也可以消除. 在图 3 (c) 中, 我们将中间的 3 组时间相同的节点合并, 再将相同节点间具有相同标识的边合并, 结果如图 3 (d) 所示. 从图 3 可以看到, 从图 3 (c) 到图 3 (d), 节点数由 8 个减少到 5 个, 边数由 10 条减少为 7 条, 格的规模减小. 实际应用中的格的规模和冗余程度都远远大于示意图中的例子, 格内合并可以很大程度地压缩格的规模, 具体数值参看实验部分. 为更进一步压缩索引规模, 我们可以放宽合并准则, 比如对有相近时间信息的节点进行聚类合并等.

5 实验

在 4 小时自然对话语音数据上通过关键词命中任务对算法进行验证. 对音素进行隐马尔科夫模型 (Hidden Markov model, HMM) 建模, 采用 39 维美尔频标倒谱系数 (Mel freguency cepstrum coefficient, MFCC) 特征, 声学模型由 154 小时朗读语音和 148 小时自然对话语音训练得到. 分别采用大词表连续语音识别器和音节识别器进行识别, 相应地采用基于词和基于音节的 Tri-gram 语言模型. 从语料标注中选取出现过的 2 至 5 字词作为目标关键词集, 共 3 979 词累计出现 5 346 次. 关键词检测任务中有两个重要指标: 虚警和召回率 (Recall, REC). 虚警指错误命中, 召回率指正确命中的关键词数占实际出现的关键词数的比例. 通过选取置信度门限调整虚警和召回率的关系, 当每小时语音的虚警个数依次为 1 到 10 时记录关键词的召回率, 其平均值称为品质因数 (Figure of merit, FOM). 选取 FOM 作为系统评价指标. 为衡量格索引的规模, 计算标注中每个字在格中对应的平均边数. 基于听写机的系统的索引规模在没有插入和删除错误时近似为 1.

5.1 基线系统与格索引

基线系统识别性能见表 1, 其中 WER、CER (Eharacter error rate)、SER (Syllable error rate) 分别表示词、字、音节的错误率. 由于词识别器的识别单元更具有区分性, 语言模型更精确且有较长的历史信息, 其识别性能最优; 音节识别器次之; 字识别器的差距是由于多音字在语言模型中的不准确建模. 表 2 中第一行 S0 给出对 LVCSR 听写机结果进行索引的检索性能, FOM 为 50.7 %. S1 是采用基于格索引的方法, FOM 提高到 69.2 %. 提升来自查全率的大幅度提高 (51.5 % ⇒ 71.2 %). 但与此同时, 索引规模也大幅度增加 (0.7 ⇒ 82.7).

5.2 不同单元识别检索性能比较

表 2 中, S1, S2, S3 和 S4 给出分别采用词、字、音节和无调音节作为识别单元和索引单元的系统性能. 可以看到, 采用音节作为识别和索引单元时, 系统有最优的检索性能 72.3 %. S1 中词格可转换为 S1.1 的字格、S1.2 的音节格和 S1.3 的无调音节格; S2 中字格转换为 S2.1 的音节格和 S2.2 的无调音节格; S3 的音节格转换为 S3.1 的无调音节格. 可以看到通过转换, 系统性能都有所提升; 其中词格转换来的无调音节格和音节格转换得到的无调音节格具有最好的性能 73.7 % 和 73.6 %.

5.3 格间融合和格内合并

表 2 中, C1 与 C2 给出格间前向融合和格内合并的结果. 通过将 1 个直接生成的无调音节格 S4 和 3 个转换得到的无调音节格 S1.3、S2.2 和 S3.1 进行后向融合, 检索性能从单个系统中最好的 73.7 % 通过格内部节点和边的合并, 格的规模被大幅度压缩到 19.3, 这个规模已经可以被应用到实际的大规模语音检索系统中. 同时, 系统性能也微弱提升 (78.6 % ⇒ 79.2 %). 这是由于格内节点和边的合并引进了新的连接关系. 如图 3 (d) 中的连接 “qing1-hua2” 在图 3 (c) 中是不存在的. 新连接增加了系统的查全率, 但也会带来大量虚警: 与 C1 相比, C2 的查全率提升要大于 FOM 提升.

表 1 识别单元性能比较 (%)

Table 1 Performance comparison of recognition units (%)

识别单元	WER	CER	SER	无调 SER
词	48.43	36.98	35.38	30.81
字	–	42.90	41.33 35.90	
音节	–	–	39.08	33.64
无调音节	–	–	–	35.83

表 2 索引性能比较 (%)

Table 2 Indexing performance comparison (%)

标号	索引类型	FOM	REC	规模
S0	听写机	50.7	51.5	0.7
S1	词格	69.2	71.2	82.7
S1.1	⇒ 字格	71.1	73.2	99.9
S1.2	⇒ 音节格	72.3	75.2	99.5
S1.3	⇒ 无调音节格	73.7	77.6	99.5
S2	字格	67.6	70.4	756.3
S2.1	⇒ 音节格	69.8	73.5	759.5
S2.2	⇒ 无调音节格	72.0	77.1	759.5
S3	音节格	72.3	76.0	110.4
S3.1	⇒ 无调音节格	73.6	78.4	110.4
S4	无调音节格	68.8	72.9	217.1
	S1.3+S2.2+S3.1+S4			
C1	格间融合	78.6	84.6	1 204.4
C2	格内合并	79.2	86.4	19.3

6 结论

本文在自然对话关键词检测任务上对汉语语音索引问题进行研究. 采用基于格的语音检索系统, 对中文中的常用识别和索引单元 — 词、字、音节和无调音节进行对比分析. 一方面在识别过程中采用不同的单元, 分别生成词格、字格、音节格和无调音节格; 另一方面, 在后验概率词格的基础上, 对不同单元的格进行转化 (词格 ⇒ 字格、音节格和无调音节格; 字格 ⇒ 音节格和无调音节格; 音节格 ⇒ 无调音节格), 以达到识别单元和索引单元的分离. 实验证明, 采用词作为识别单元、无调音节作为索引单元具有最好的检索性能 (FOM = 73.7 %), 采用音节和无调音节分别作为识别单元和索引单元也可达到基本

相当的检索性能 (FOM = 73.6 %). 通过不同系统生成的格间的融合, 系统性能得到大幅度提高 (FOM: 73.7 % ⇒ 78.6 %). 融合格的规模较大, 通过对其内部节点和边的合并, 格规模大大降低至 20 以内, 为大规模语音数据库的索引提供了可能.

References

1 Information Access Division. Spoken document retrieval evaluation [Online], available: http://www.itl.nist.gov/iad/mig/tests/sdr, December 14, 2009

2 Saraclar M, Sproat R. Lattice-based search for spoken utterance retrieval. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Boston, USA: Association for Computational Linguistics, 2004. 129−136

3 Yu P, Chen K, Ma C, Seide F. Vocabulary-independent indexing of spontaneous speech. *IEEE Transactions on Speech and Audio Processing*, 2005, **13**(5): 635−643

4 Zhou Z Y, Yu P, Chelba C, Seide F. Towards spoken document retrieval for the internet: lattice indexing for large-scale web-search architectures. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. New York, USA: Association for Computational Linguistics, 2006. 415−422

5 Mamou J, Ramabhadran B, Siohan O. Vocabulary independent spoken term detection. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam, The Netherlands: ACM, 2007. 615−622

6 Ng K, Zue V W. Subword-based approaches for spoken document retrieval. *Speech Communication*, 2000, **32**(3): 157−186

7 Wallace R G, Vogt R J, Sridharan S. A phonetic search approach to the 2006 NIST spoken term detection evaluation. In: Proceedings of the 8th Annual Conference of the International Speech Communication Association. Antwerp, Belgium: NIST, 2007. 2385−2388

8 Lee S W, Tanaka K, Itoh Y. Combining multiple subword representations for open-vocabulary spoken document retrieval. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Philadelphia, USA: IEEE, 2005. 505−508

9 Yu P, Seide F. A hybrid word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech. In: Proceedings of International Conference on Spoken Language Processing. Jeju Island, Korea: IEEE, 2004. 293−296

10 Chen B, Wang H M, Lee L S. Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characters. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Istanbul, Turkey: IEEE, 2000. 1771−1774

11 Wang H M, Meng H, Schone P, Chen B, Lo W K. Multi-scale audio indexing for translingual spoken document retrieval. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Salt Lake City, USA: IEEE, 2001. 605−608

12 Meng Meng, Wang Xiao-Rui, Liang Jia-En, Xu Bo. A system combination based keyword-spotting method using complementary acoustic models. *Acta Automatica Sinica*, 2009, **35**(1): 39−45
(孟猛, 王晓瑞, 梁家恩, 徐波. 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. 自动化学报, 2009, **35**(1): 39−45)

13 Tu Xin-Hui. A Study of Some Issues in Chinese Text Information Retrieval [Master dissertation], Huazhong Normal University, China, 2006
(涂新辉. 中文文本信息检索相关技术研究 [硕士学位论文], 华中师范大学, 中国, 2006)

14 Meng Sha, Yu Peng, Seide F, Liu Jia. Indexing of posterior lattice for spontaneous Mandarin speech. *Journal of Tsinghua University (Science and Technology)*, 2008, **48**(z1): 673−677
(孟莎, 余鹏, Frank Seide, 刘加. 基于后验概率词格的汉语自然对话语音索引. 清华大学学报 (自然科学版), 2008 **48**(z1): 673−677)

15 Wessel F, Schluter R, Macherey K, Ney H. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2001, **9**(3): 288−298

16 Meng S, Yu P, Seide F, Liu J. A study of lattice-based spoken term detection for Chinese spontaneous speech. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding. Kyoto, Japan: IEEE, 2007. 635−640

孟 莎 清华大学电子工程系博士研究生. 2004 年获华中科技大学电子与信息工程系通信工程学士学位. 主要研究方向为语音识别与语音检索.
E-mail: mengs04@mails.thu.edu.cn
(**MENG Sha** Ph. D. candidate in the Department of Electronic Engineering, Tsinghua University. She received her bachelor degree from Huazhong University of Science and Technology in 2004. Her research interest covers speech recognition and speech retrieval.)

余 鹏 微软亚洲研究院研究员. 2002 年获清华大学电子工程系博士学位, 主要研究方向为语音识别与语音检索.
E-mail: rogeryu@microsoft.com
(**YU Peng** Researcher at Microsoft Research Asia. He received his Ph. D. degree from Tsinghua University in 2002. His research interest covers speech recognition and speech retrieval.)

刘 加 清华大学电子工程系教授. 主要研究方向为语音识别、说话人识别、语种识别与语音芯片设计. 本文通信作者.
E-mail: liuj@tsinghua.edu.cn
(**LIU Jia** Professor in the Department of Electronic Engineering, Tsinghua University. His research interest covers speech recognition, speaker recognition, language identification and speech processing chip design. Corresponding author of this paper.)